Assessing the Quality of Uncertainty Calibration

Katherine E. Brown, PhD[1,2], Steve Talbert, PhD[3], Douglas A. Talbert, PhD[1]
[1]Tennessee Tech University, Cookeville, TN; [2]Vanderbilt University Medical Center, Nashville, TN; [3]University of Central Florida, Orlando, FL

Introduction

Uncertainty quantification provides an indicator of potential incorrectness in opaque machine learning models, and standard empirical evaluations of the efficacy of uncertainty typically involves generating and visually inspecting rejection-classification plots[1,2]. Such plots measure the change in a desired unit of measurement as the most uncertain data are removed from consideration. The desired effect is a steep increase as the most uncertain data are removed and then for performance to plateau. To perform a more formal, statistical analysis of rejection-classification plots requires converting these to numeric representation, which can be accomplished by taking the area under the rejection-classification plot (AURCP). We argue, however, that this value is only useful in comparison to other AURCP values using the same dataset and machine learning model. This is because the rejection-classification plots do not a have standard starting point like ROC curves. The baseline accuracy for a model applied to a complete dataset set (prior to the removal of any uncertain data) provides a non-zero starting point that prevents a single AURCP value from revealing anything about the quality of uncertainty calibration. The meaning of an AURCP value is completely dependent on the model's baseline accuracy, and thus, meaningless as a single number. This poster describes and demonstrates a novel adaptation of AURCP that addresses this limitation.

Table 1. RC-Index values for the trauma triage data. Note higher values are better.

Bayesian	Softmax	Control
0.084	0.076	-0.0007

Empirical Analysis and Results

We used data from a Level 1 Trauma Center (Trauma) that included physiological parameters, anatomical criteria, mechanism of injury, and age with severe injury defined by an injury severity score > 15. We train a deep neural network and collect uncertainty of predictions using Bayesian dropout[3] and using the output probabilities (defined as $u(x) = 1/|p(x) - 0.5|$ where *p(x)* represents the model's predicted probability of the positive class). We produce rejection-classification curves based on these uncertainty values by removing the most uncertain predictions in 5% increments and remeasuring accuracy. The resulting curves are compared to a control that removes data randomly in the same increments. In addition to presenting the Rejection-Classification curves, we also present the Rejection-Classification Index (RC-Index), our novel measure based on numerical integration under the rejection-classification curve when normalized as follows: The accuracy over the complete dataset is subtracted from all other accuracy values. The possible range of values for RC-Index is -1 to 1. The sign of RC-Index indicates positive or negative change in accuracy as more uncertain data are removed. All reported values are averaged over 10-fold stratified cross validation.

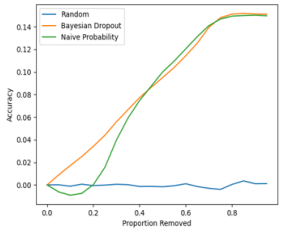

Figure 1. Rejection Classification Curve.

Discussion and Conclusions

Figure 1 presents the rejection-classification plots for the trauma triage data. We can see that both measures of uncertainty produce a usable rejection-classification compared to the random control. It is evident that Bayesian dropout produces a more consistent, monotonic increase in accuracy than using the uncertainty measure derived from the network's softmax output which decreases in performance before increasing to the same plateau as Bayesian dropout. This is reflected in the RC-Index values for these uncertainty measures, with Bayesian dropout having an RC Index 0.008 higher than RC-Index of the softmax probability. Most importantly, each RC-Index value is informative on its own, with both the Bayesian and softmax values indicating some degree of correct calibration and the control value indicating no calibration at all. Thus, RC-Index is an improvement over AURCP because it provides value both as a standalone metrics and for comparing different approaches to uncertainty calibration. Future work will include tuning this measure to better account for non-monotonic performance increases.

Acknowledgements. This publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R15LM013824.

References

1. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports. 2017 Dec 19;7(1):1-4.
2. Brown KE, Talbert DA. Estimating Uncertainty in Deep Image Classification. In: Proceedings of the American Medical Informatics Association Annual Symposium. 2019.
3. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. 2016 Jun 11 (pp. 1050-1059). PMLR.