Selective Classification for Learning-enabled Cyber-Physical Systems based on Sequential Data

Xenofon Koutsoukos and Dimitrios Boursinos
Institute for Software Integrated Systems, Vanderbilt University

ABSTRACT

Cyber-Physical Systems (CPS) integrate computing, monitoring and control for operation in the physical world. Perception of the environment is a complex process because of the existence of objects that are difficult to model and have complex interaction with the controlled system. Deep Neural Networks (DNNs) have the capacity to be trained and generalize their knowledge to make predictions in dynamic environments. CPS can benefit from the integration of DNNs but assurance guarantees are needed that are very challenging to compute. In CPS applications such as autonomous vehicles that need to perceive the environment and take the correct control decisions, the cost of an incorrect classification is much higher than not making any classification when there is no clear distinction between the best prediction and the alternatives. In such a setting, the operation cost over time can be minimized using selective classifiers that evaluate the risk in each classification and either accept the classification or reject it.

Most discriminative machine learning (ML) frameworks make predictions with some notion of confidence under the assumption that the input data are independent and identically distributed (IID). However, when this assumption does not hold the confidence metrics are not accurate. This is an important challenge for making decision in learning-enabled autonomous systems. Sensors perceive phenomena in the physical environment that have some duration and data from individual time instances have some, usually unknown, dependence to previous instances. This leads to miscalibration of confidence estimates and the error-rate of ML algorithms obtained during design time are not satisfied during the system operation.

Our approach for improving the confidence of the predictions is based on Inductive Conformal Prediction (ICP). ICP aims in producing prediction sets that satisfy any error-rate bound guarantees under the IID assumption. The main idea is to test if a new input example conforms to the training data set by utilizing a nonconformity measure which assigns a numerical score indicating how different the input example is from the training data set. For any test input, a p-value is assigned to each possible class to decide if a class should be part of the prediction set or not in order to satisfy the chosen error-rate guarantees. ICP provides well-calibrated confidence measure for IID data.

The goal of this work is to improve the calibration of the prediction sets computed by ICP and the classification accuracy when the input data are time correlated. We use statistical methods for computing aggregated p-values resulted from sequential data. We approach the problem as a multiple hypothesis testing problem and show how different combination methods recover ICP validity. Our main contribution is the design of a selective classifier based on the aggregate p-values for each class. When the highest p-value among all the classes is much higher than the second highest computed p-value we can trust the classification more than cases where at least two of the highest p-values are close to each other. Another contribution of this work is the computation of low-dimensional, appropriate, embedding representations of the original inputs in a space where the Euclidean distance is a measure of similarity between the original inputs. This is needed in order to find semantic similarities between data points and handle high-dimensional inputs in real-time. We evaluate the presented approaches on the German Traffic Sign Recognition Benchmark (GTSRB) which provides sequential images of signs as a vehicle moves towards traffic signs.