Anytime Prediction: Efficient Ensemble Methods for Any Computational Budget

Alex Grubb

Thursday, June 21, 2012
3:00 p.m.
GHC 7501

A modern practitioner of machine learning must often consider trade-offs between accuracy and complexity when selecting from available machine learning algorithms. Prediction tasks can range from requiring real-time performance, as in web applications which must handle thousands of requests per second with little latency, to being largely unconstrained in their use of computational resources. In each setting, an ideal algorithm utilizes as much of the available computation as possible to provide the most accurate result.

This issue is further complicated by applications where the computational constraints are not fixed in advance. In robotics applications predictions are often needed in time to allow for adaptive behaviors which respond to real-time events. Such constraints often rely on a number of factors at prediction time, making it difficult to select a fixed prediction algorithm a priori. In these situations, an ideal approach is to use an anytime prediction algorithm. Such an algorithm rapidly produces an initial prediction and then continues to refine the result as time allows, producing final results which dynamically improve to fit any computational budget.

This thesis proposes a framework for learning and analyzing anytime predictors built from ensembles of weaker predictors. Our approach incrementally applies a sequence of weaker predictors as time progresses, using each new result to update and improve the current prediction as time allows. By greedily selecting weak predictors which obtain the best performance increase per unit of computational cost our algorithms provide an intuitive and effective way to trade-off computational cost and accuracy for any computational budget.

We present the results of applying our algorithms to a number of problem domains such as classification and object detection that indicate that our approach to anytime prediction is more efficient than trying to adapt a number of existing methods to the anytime prediction problem. We further explore how the incremental nature of our algorithms can be used to build weak predictors for structured prediction which focus computational resources on parts of the problem where they would be most effective. We also examine extensions of this framework designed to handle parallel resource utilization and feature acquisition costs.

Thesis Committee:
J. Andrew Bagnell, Chair
Avrim Blum
Martial Hebert
Alexander Smola, Google Research