IEEE Conference on Evolving and Adaptive Intelligent Systems
June 2 — 4, 2014
Linz, Austria

Abstract:

SEMI-SUPERVISED AND ACTIVE LEARNING IN INITIALLY LABELED NONSTATIONARY AND EVOLVING ENVIRONMENTS

Author: Robi Polikar

An increasing number of real-world applications are associated with streaming data drawn from drifting and nonstationary distributions that change over time. These applications demand new algorithms that can learn and adapt to such changes, also known as concept drift. Proper characterization of such data with existing approaches typically requires substantial amount of labeled instances, which may be difficult, expensive or even impractical to obtain. Such a scenario is also related to the problem known as *verification latency*, where the labels of the training data are not available until much later than the data itself – or in *extreme verification latency* that we discuss in this talk – they may never be available. In the first half of this lecture, we will introduce COMPOSE, a density tracking framework for learning from nonstationary streaming data, where labels are unavailable (or presented very sporadically) after initialization. We will discuss the algorithm in detail, as well as its results and performances on real-world datasets as well as several carefully constructed synthetic datasets, which demonstrate the ability of the algorithm to learn under several different scenarios of *initially labeled streaming environments* (ILSE). Furthermore, we also demonstrate that COMPOSE is competitive even with a well-established, fully supervised, nonstationary learning algorithms that receive labeled data in every batch. COMPOSE, like all algorithms, make certain assumptions on the data distribution, the most important of which is the “limited-drift” assumption, where it assumes that any class distribution at two consecutive time-steps have significant overlap, i.e., the drift is gradual. In addition to such cases as abrupt drift, COMPOSE also cannot address special cases such as introduction of a new class or significant overlap *among* existing classes, as such scenarios cannot be learned without additional labeled data. Scenarios that provide occasional or periodic limited labeled data are not uncommon, however, for which many of COMPOSE’s restrictions can be lifted. In the second part of this talk, I will briefly introduce an alternate version of COMPOSE as a proof-of-concept algorithm that can identify the instances whose labels – if available – would be most beneficial, and then combine those instances with unlabeled data to actively learn from streaming nonstationary data, even when the distribution of the data experiences abrupt changes.