Ph.D. in Information Technology
Thesis Defense

February 29th, 2024
at 10:00 am
Sala Conferenze Emilio Gatti

Diego STUCCHI – XXXV Cycle

ALGORITHMS FOR THE MULTIVARIATE CHANGE DETECTION. EXTENDING THE QUANTTREE

Supervisor: Prof. Giacomo Boracchi

Abstract:

Change detection has been a relevant research topic for decades, and change detection algorithms find application in numerous real-world domains, from industrial monitoring to healthcare, from security to finance. In practice, change detection problems consist of monitoring a stream of independent and identically distributed (i.i.d.) samples that initially follow a stationary distribution. The goal of a change detection algorithm is to detect when the process generating the stream drifts towards a different unknown post-change distribution. Indeed, the timely detection of distribution changes is crucial in real-world applications, as these changes might indicate faults in the monitored machinery or changes in the operating conditions and might require measures to counteract the drift.

In this thesis, I solve the change detection problem by nonparametric algorithms, namely, assuming that both the stationary and post-change distributions are unknown.

Depending on the application, the problem of detecting a change in a stream can be addressed in two manners: the batch-wise setting, where data are analyzed in fixed-size windows, and the online monitoring, where data are processed as a stream of continuous samples and are often associated with labels. In my research, I tackled both batch-wise and online monitoring problems.

The main focus of my research on change detection has been extending the QuantTree algorithm, a solution based on a partitioning of the input space and supported by sound theoretical results. The main contributions of my work are the Kernel QuantTree (KQT) algorithm and the MultiModal QuantTree, each solving a specific limitation of QuantTree while generalizing its theoretical results. The histogram construction by QuantTree employs axis-aligned splits of the input space. For KQT, I designed a novel splitting rule based on measurable kernel functions, which produce finite-volume bins that better adhere to the data distribution. Moreover, I proved that the control of the FPR generalizes to this histogram construction algorithm. In particular, I have proposed three measurable quadratic functions and proved that the resulting KQT is invariant under roto-translations. Thanks to this property, KQT does not require classical preprocessing methods -- like PCA -- to achieve remarkable detection performance.

Another limitation of QuantTree, common to all change detection algorithms, is the assumption that stationary data follow a single distribution. Such an assumption does not match many real-world scenarios where, in stationary conditions, the monitored process alternates between different normal operating settings, i.e., modalities. I formalized this multimodal change detection problem and solved it by MMQT, which uses a single modality-agnostic histogram to characterize

the stationary distributions and computes modality-specific statistics for detecting changes. For the MMQT algorithm, I leveraged the theoretical properties of QuantTree to i) automatically estimate the number of modalities in a training set and ii) derive a principled calibration procedure that guarantees false-positive control.
On top of these contributions to the batch-wise change detection, I also tackled the supervised online monitoring problem. The concept-drift detection problem is often associated with a classification task, and many solutions monitor the error rate of a classifier and detect distribution changes that harm the classification performance but ignore drifts that have little impact on the error rate, which are called virtual drifts. Other solutions monitor the stream distribution to detect any deviation from the stationary distribution, independently of the impact on related classification problems but ignore the supervision granted by the labels.
To combine the supervised information with the distribution monitoring, I designed the Class-Distribution Monitoring (CDM) algorithm, which monitors the class-conditional distributions of a datastream by leveraging multiple instances of QT-EWMA -- an online monitoring algorithm based on QuantTree. CDM reports a concept drift after detecting a distribution change in any class, thus identifying which classes are affected by the concept drift. This information is precious for diagnostics and adaptation techniques, which can update only the part affected by the drift.

PhD Committee

Prof. **Daniele Loiacono, Politecnico di Milano**

Prof. **Alessandro Giusti, IDSIA USI-SUPSI**

Dr. **Cristiano Cervellera, CNR**