DLI Accelerated Data Science Teaching Kit

Lecture 18.4 - Reasoning and Data Resource

Monitor Learning System

Learning from data streams is a continuous process.

The learning systems need to monitor their working conditions since they act in dynamic environments, where working conditions change and evolve.

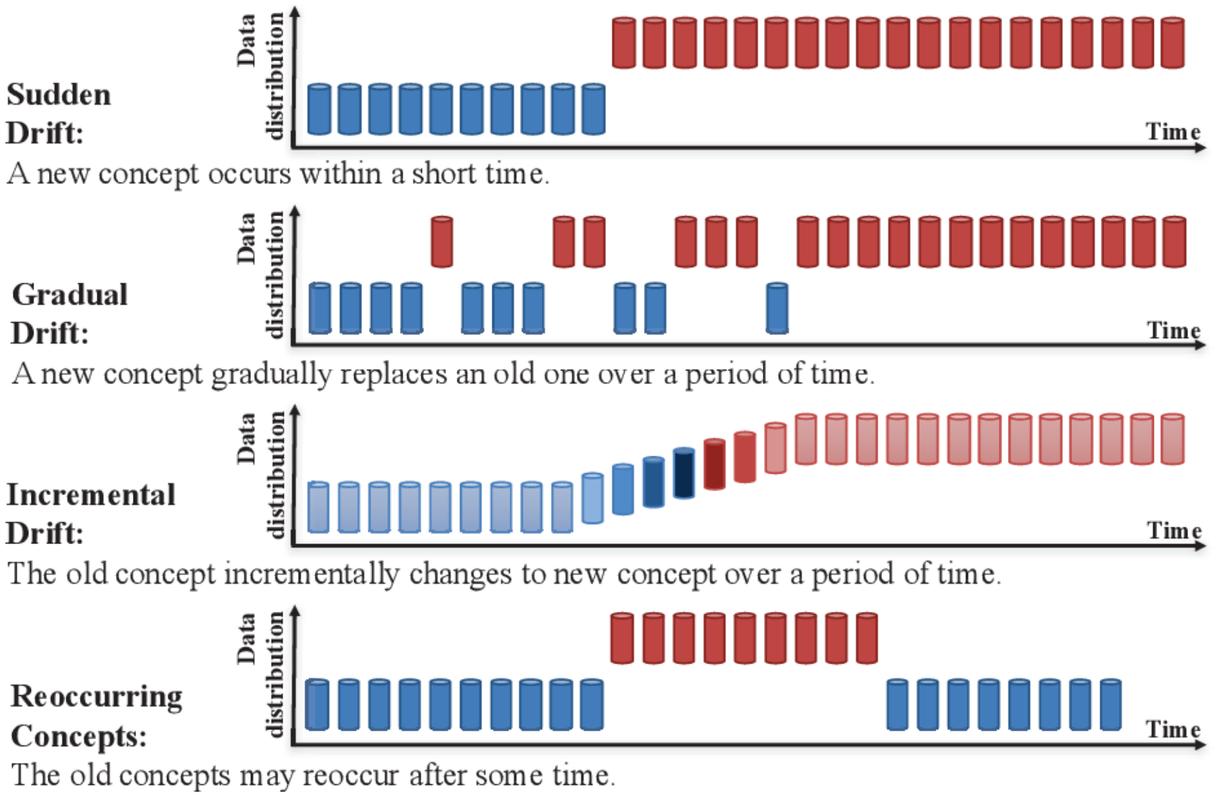
They need to monitor the learning process for change detection, emergence of novel classes, changes in the relevance of features, changes in the optimal parameter settings, and others.

In a general sense, these learning systems should be able of self-diagnosis when performance degrades, by identifying the possible causes of degradation and self-repairing or self-reconfiguring to recover to a stable status.

Concept Drift

During classification, a change in the concept or distribution of dataset over the time is termed as concept drift.

Algorithms focusing on drift detection on delayed or partially labeled streams exists.



Source: https://www.semanticscholar.org/paper/Concept-drift-adaptation-learning-with-data-Liu/5b105e357936f989cfb46ddd055ea44a2b0aed04

Feature Drift

Data streams are subject to different types of feature drifts.

- changes in the values of a feature and their association with the class
- changes in the domain of features
- changes in the subset of features that are used to label an instance

A feature drift occurs when a subset of features becomes relevant to the learning task.

The assessment of feature drifting scenarios should not only account for the accuracy rates of learners, but also whether the feature selection process correctly flags the changes in the relevant subset of features and if it identifies the features.

Hyperparameter Tuning for Evolving Data Streams

Hyperparameter tuning (or optimization) is often treated as a manual task where experienced users define a subset of hyperparameters and their corresponding range of possible values to be tested exhaustively.

The brute force approach of trying all possible combinations of hyperparameters and their values is time-consuming but can be efficiently executed in parallel.

The challenge is to design an approach that incorporate the hyperparameter tuning as part of the continual learning process, which might involve data preprocessing, drift detection, drift recovery, and others.

Data Resources

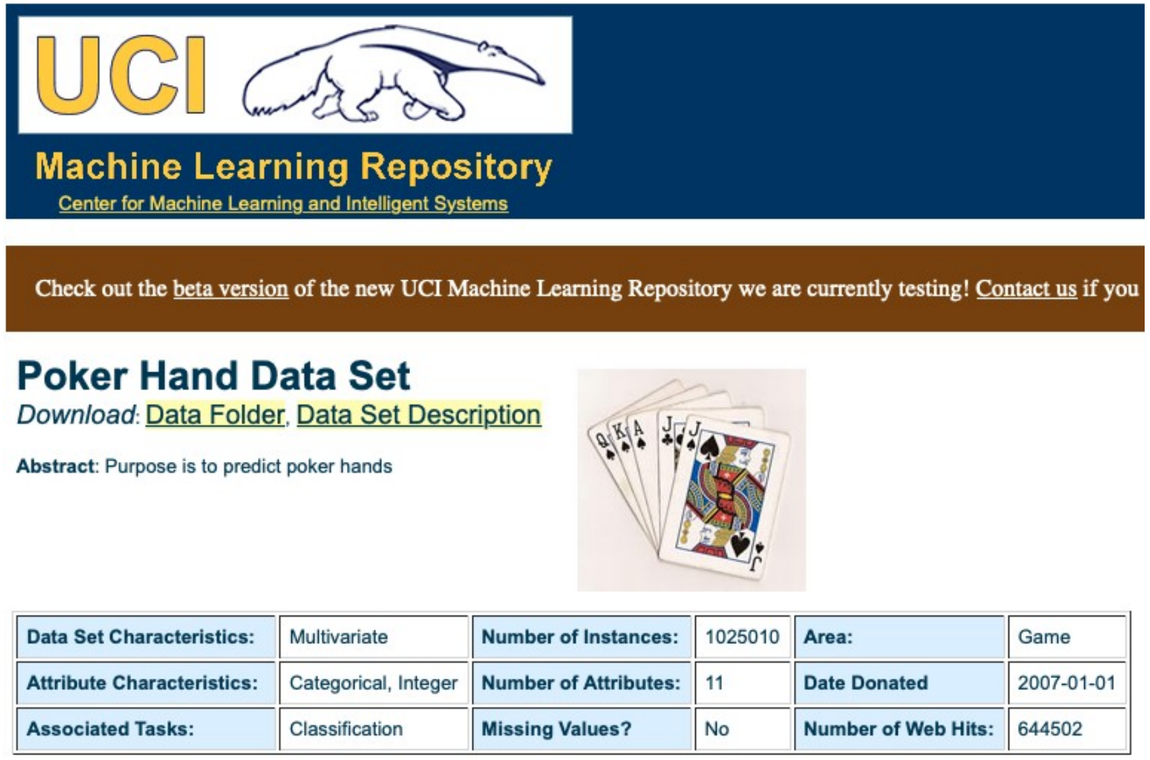
Data stream algorithms are usually assessed using a benchmark that is a combination of synthetic generators and real-world datasets.

The synthetic data is used to allow showing how the method performs given specific problems (e.g., concept drifts, concept evolution, feature drifts, and so forth) in a controlled environment.

The real-world datasets are used to justify the application of the method beyond hypothetical situations; however, they are often used without guarantees that the issues addressed by the algorithm are present. For example, it is difficult to check if, and when, a concept drift takes place in a real dataset.

Data Resources

When it comes to real-world data streams, some researchers use datasets that do not represent data streams or that are synthetic data masquerade as real datasets.



Source: https://archive.ics.uci.edu/ml/datasets/Poker-

DLI Accelerated Data Science Teaching Kit

Thank You