Part 6 - Stream Mining

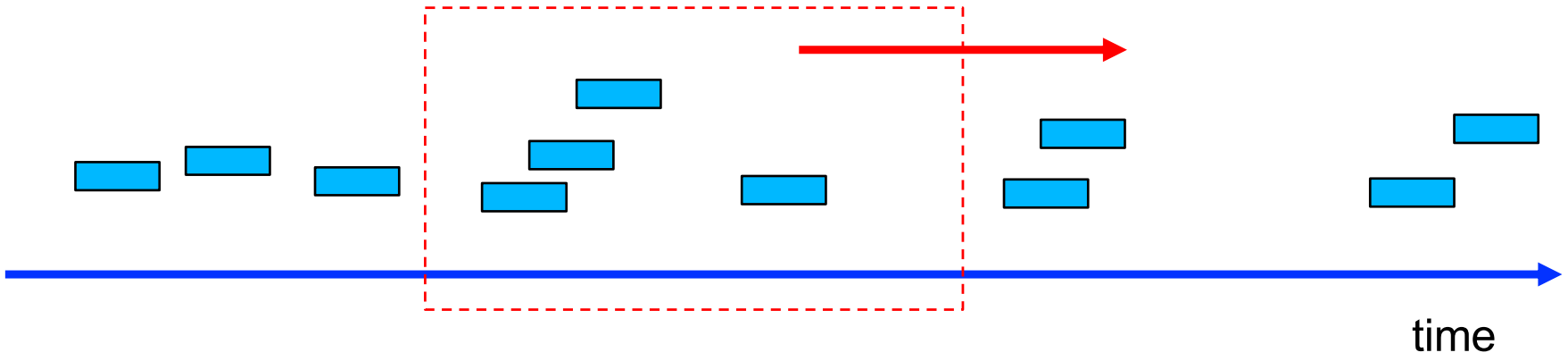
Adapted from:

DATA STREAM MINING: A Practical Approach. By Albert Bifet and Richard Kirkby, 2009.

Agenda

- Introduction
- Data Stream Mining Basics
  - Requirements
  - Cycle
  - Challenges

Data streams



Introduction

- Data mining allows larger data sets to be handled;
- Data doesn’t fit memory?
  - Sample, temporary storage, subsets…
- New data is available?
  - Retrain the model with the addition of new data;

INEFFICIENT

Introduction

- Solution: Data Stream Mining
  - Naturally copes with data sizes many times greater than memory;
  - Works with challenging **real-time applications** (difficult for ML or DM);
  - Training examples arrive in a **high speed, potentially endless, stream**;
  - **Updates its model incrementally** as each example is inspected;
  - Capable of giving a prediction at any time;

Agenda

- Introduction
- Data Stream Mining Basics
  - Requirements
  - Cycle
  - Challenges

Data Stream Mining Basics

- Data Stream Requirements:
  1) Inspect each example only once;
  2) Use limited amount of memory;
  3) Work in a limited amount of time;
  4) Be ready to predict at any point.

Data Stream Mining Basics

- Classification cycle:

1) Receives example;
2) Updates data structures;
3) Ready for the next example or prediction.

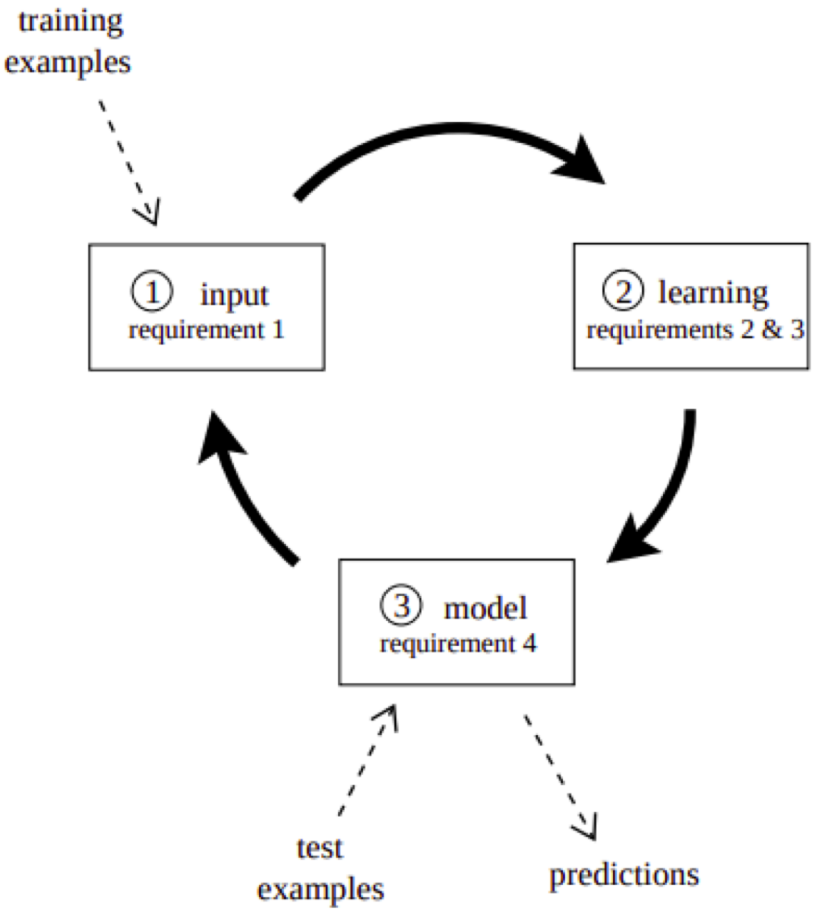


Figure 1.1: The data stream classification cycle.

Data Stream Mining Basics

- Possible approaches:
  1) Wrapper – creates batches of examples to be applied using traditional batch learner to induce a model.
     Challenges:
     - Determining appropriate batch sizes (size VS response time VS quality);
     - Training times not controlable by a wrapper algorithm;
     - Memory management.
  2) Adaptation - adapted algorithms designed specifically for data stream problems. Advantages:
     - Greater control over processing time per example;
     - Better memory management at a finer-grained level.

Data Stream Mining Basics

- Challenges
  1) Streams are not static and can change over time;
  2) Harder to evaluate performance;

Data Stream Mining Basics

- Change detection
  1) Concept Change
     - Concept drift
     - Concept shift
  2) Distribution or sampling change

Data Stream Mining Basics

- Concept is the class being predicted.
  - **Concept change** is change of underlying concept over time;
    - **Drift** describes a gradual change.
    - **Shift** describes an abrupt change.
  - **Distribution change** refers to data distribution change. It can happen while the concept stays the same. Increases the error rate of the model.
- The design of a change detector is a tradeoff between detecting true changes and avoiding false alarms.

Data Stream Mining Basics

- Example of strategies for detecting changes
  - CUSUM Test
  - Geometric Moving Average (GMA)
  - Statistical Test
  - Drift Detection Method (DDM)
  - Exponential Weighted Moving Average (EWMA)

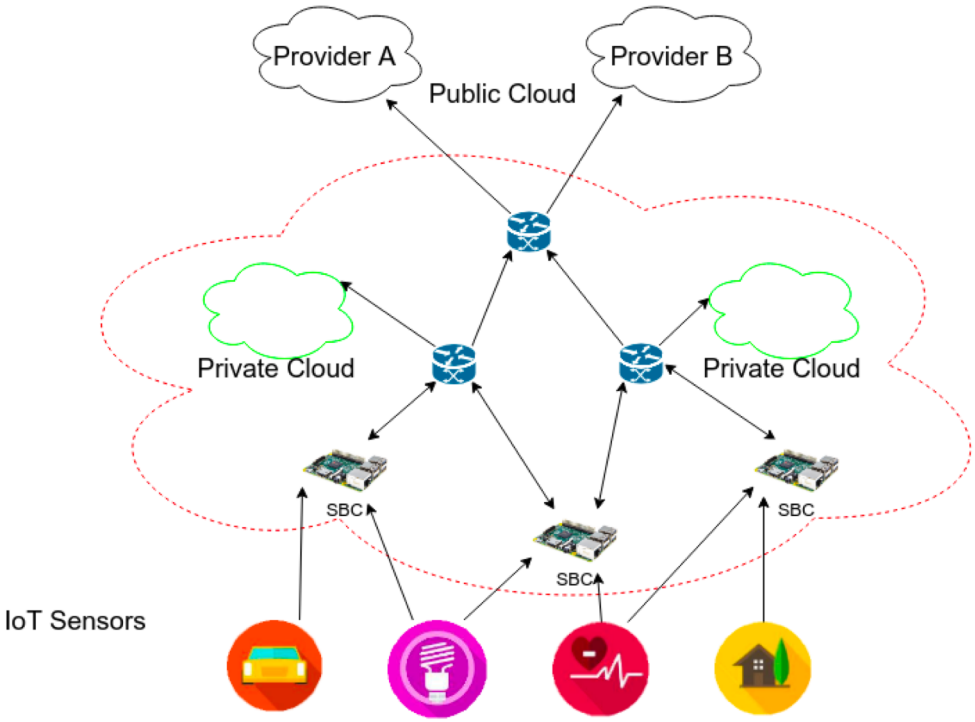
Data Stream Mining Basics

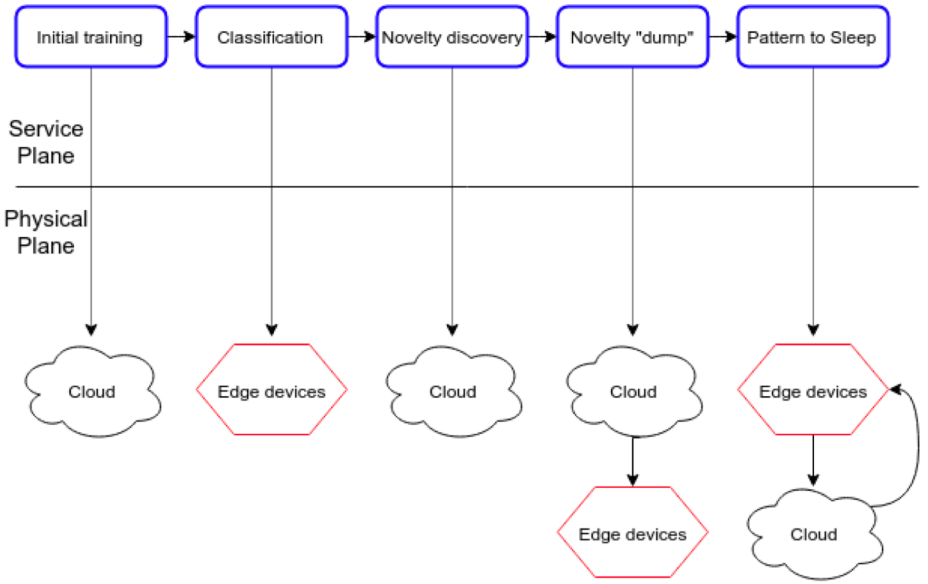
- Challenges
  1) Streams are not static and can change over time;
  2) Harder to evaluate performance;

Data Stream Mining Basics

- Performance Evaluation
  - Holdout: alternative to CV, creates a train subset and a test subset;
  - Interleaved Test-Then-Train: does both for each example arriving;

Example: IDSA-IoT

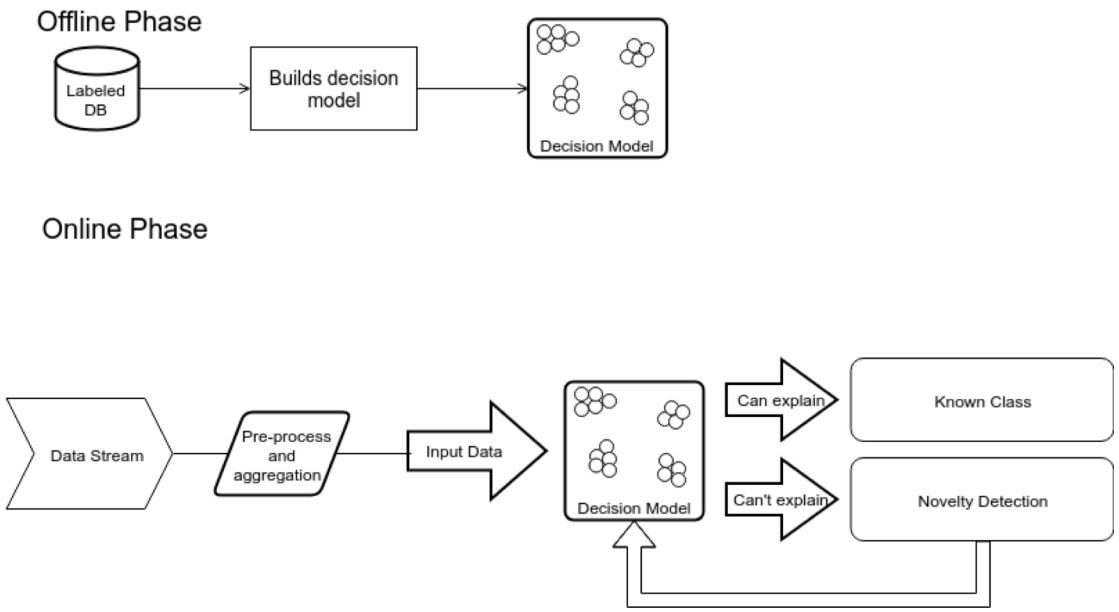




Cassales, G. W., Senger, H., de Faria, E. R., & Bifet, A. (2019, June). ***IDSA-IoT: An Intrusion Detection System Architecture for IoT Networks***. In *2019 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1-7). IEEE.

Evaluation

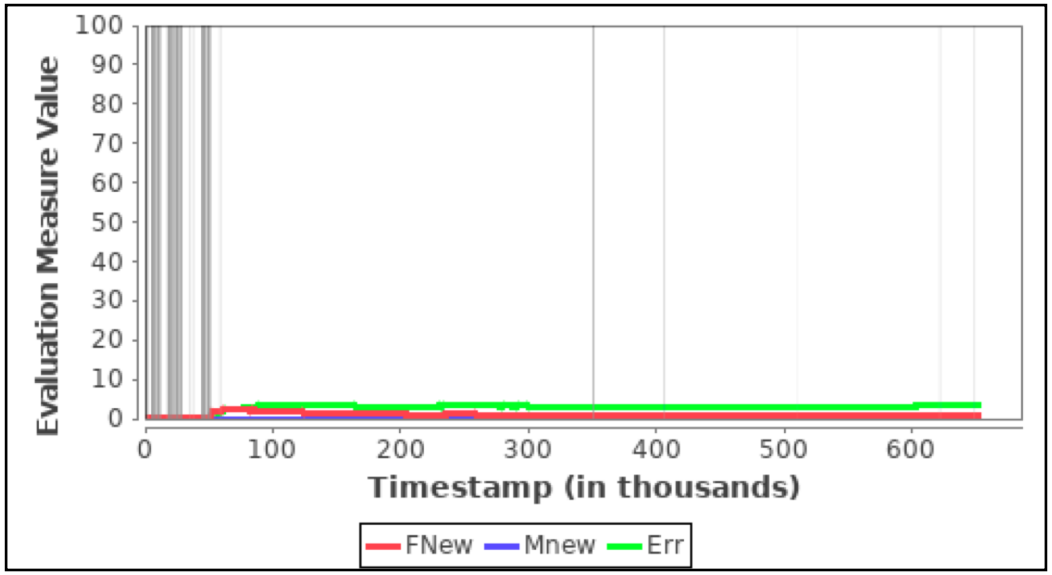
- Three algorithms
  - MINAS
  - ECSMiner
  - AnyNovel
- Dataset
  - Kyoto



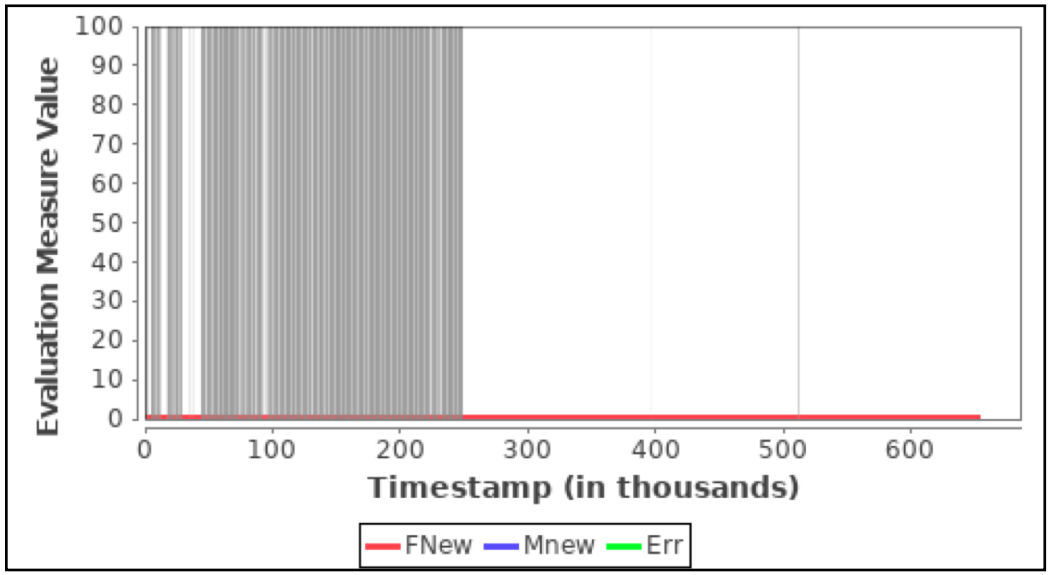
Metrics

- Fnew: the % of normal class examples incorrectly identified as novelty
- Mnew: the % of novelty examples erroneously classified as normal class
- Error: (FN + FP) / number of examples classified

Results



Binary classification with a label delivery delay of 50,000 examples and total feedback (ECSMiner).



Binary classification with label delivery delay of 50,000 examples and partial feedback of 1% (ECSMiner)

Exercise

- Install MOA
  - https://moa.cms.waikato.ac.nz/getting-started/
  - Unzip the file
- Windows systems:
  - Execute moa-release-2019.05.0/bin/moa.bat
- Linux/MacOS:
  - Execute moa-release-2019.05.0/bin/moa.sh

•Go to directory: part-6-Stream-Mining/Lab

•Open file 1.LAB-MOA-Intro