MOA: {M}assive {O}nline {A}nalysis.

Albert Bifet

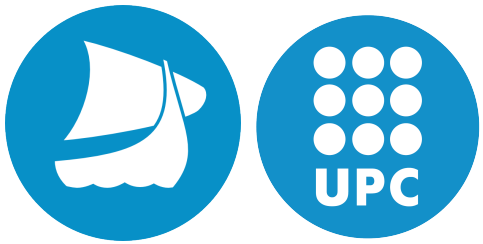
THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Hamilton, New Zealand

August 2010, Eindhoven

Adaptive Learning and Mining for Data Streams and Frequent Patterns

Coadvisors: Ricard Gavaldà and José L. Balcázar

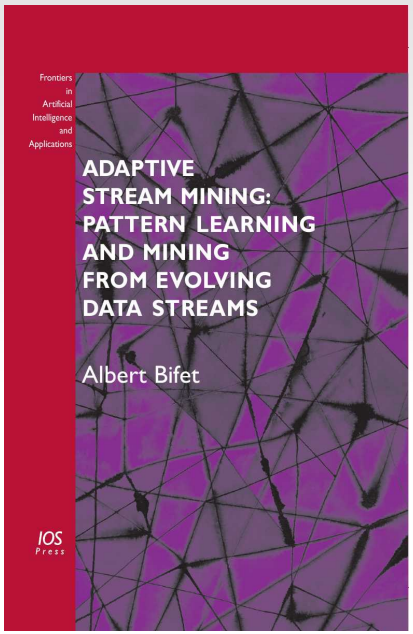


LARCA

Laboratory for Relational Algorithmics, Complexity and Learning

Universitat Politècnica de Catalunya

Adaptive Stream Mining



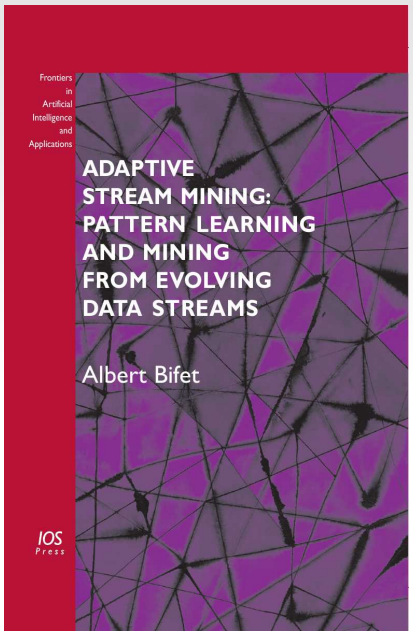
Mining Evolving Streams

1. Framework
2. ADWIN
3. Hoeffding Adaptive Tree
4. ADWIN Bagging
5. Adaptive-Size Hoeffding Tree Bagging

Mining XML Tree Streams

6. Closure Operator on Trees
7. Incremental Method
8. Sliding Window Method
9. Adaptive Method
10. XML Classification

Adaptive Stream Mining



Mining Evolving Streams

1. Framework
2. `ADWIN`
3. Hoeffding Adaptive Tree
4. `ADWIN` Bagging
5. Adaptive-Size Hoeffding Tree Bagging

Mining XML Tree Streams

6. Closure Operator on Trees
7. Incremental Method
8. Sliding Window Method
9. Adaptive Method
10. XML Classification

Mining Massive Data

2007

- Digital Universe: 281 exabytes (billion gigabytes)
- The amount of information created exceeded available storage for the first time

Web 2.0

- 106 million registered users
- 600 million search queries per day
- 3 billion requests a day via its API.

Green Computing

Green Computing

Study and practice of using computing resources efficiently.

Algorithmic Efficiency

A main approach of Green Computing

Data Streams

Fast methods without storing all dataset in memory

What is MOA?

{M}assive {O}nline {A}nalysis is a framework for online learning from data streams.

- It is closely related to WEKA
- It includes a collection of offline and online methods as well as tools for evaluation:
  - boosting and bagging
  - Hoeffding Trees

  with and without Naïve Bayes classifiers at the leaves.

What is MOA?

- Easy to extend
- Easy to design and run experiments

Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl, Albert Bifet, Geoff Holmes, Bernhard Pfahringer

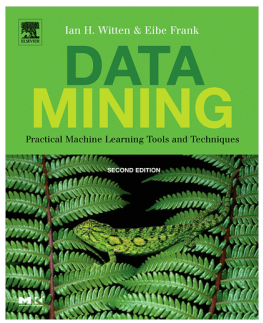
RWTH Aachen University, University of Waikato

Benchmarking Stream Clustering Algorithms within the MOA Framework

KDD 2010 Demo

WEKA

- Waikato Environment for Knowledge Analysis
- Collection of state-of-the-art machine learning algorithms and data processing tools implemented in Java
  - Released under the GPL
- Support for the whole process of experimental data mining
  - Preparation of input data
  - Statistical evaluation of learning schemes
  - Visualization of input data and the result of learning

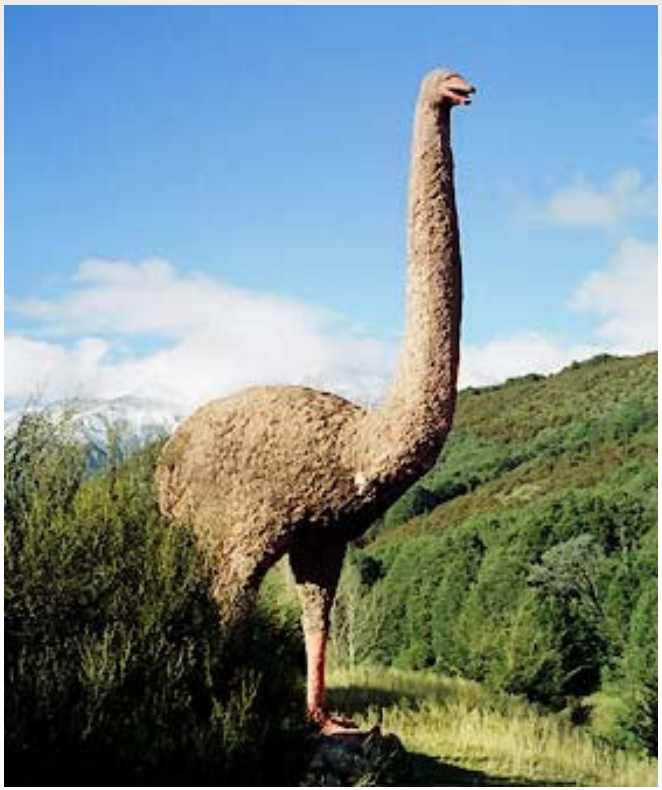


- Used for education, research and applications
- Complements “Data Mining” by Witten & Frank

WEKA: the bird

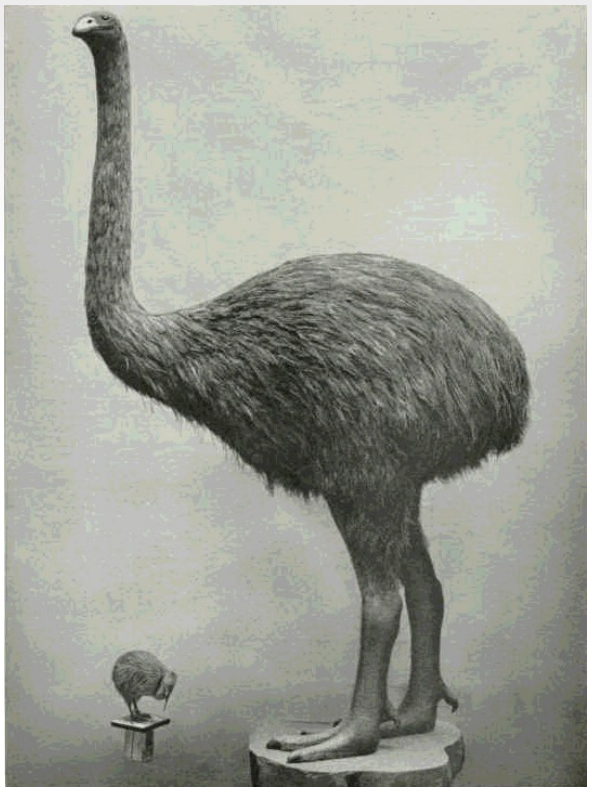
MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.



MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

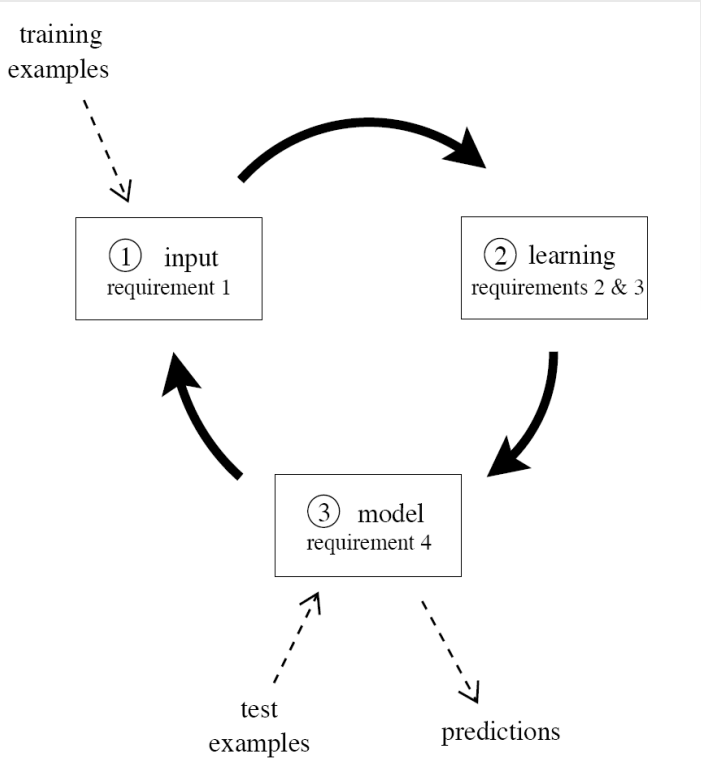


MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

Data stream classification cycle

1. Process an example at a time, and inspect it only once (at most)
2. Use a limited amount of memory
3. Work in a limited amount of time
4. Be ready to predict at any point



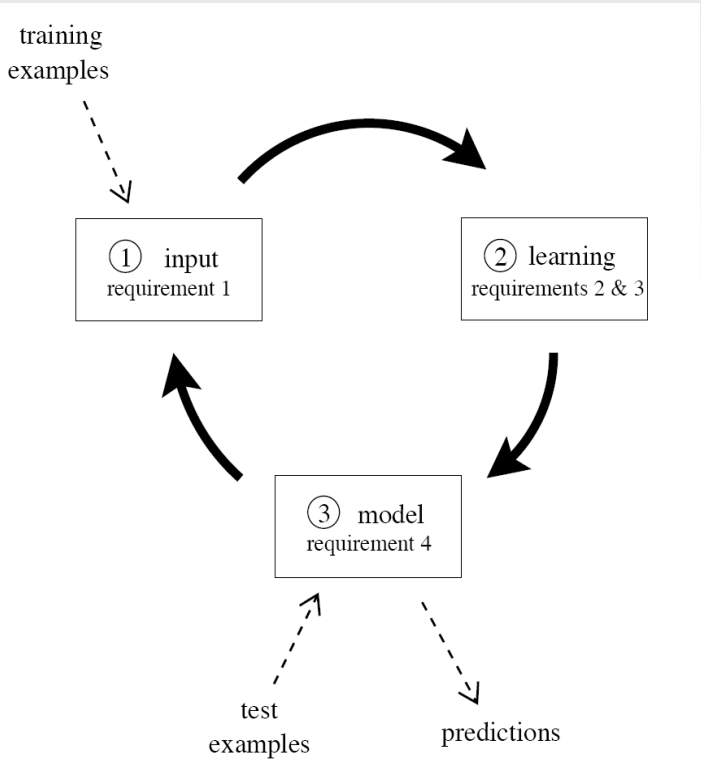
Experimental setting

Evaluation procedures for Data Streams

- Holdout
- Interleaved Test-Then-Train or Prequential

Environments

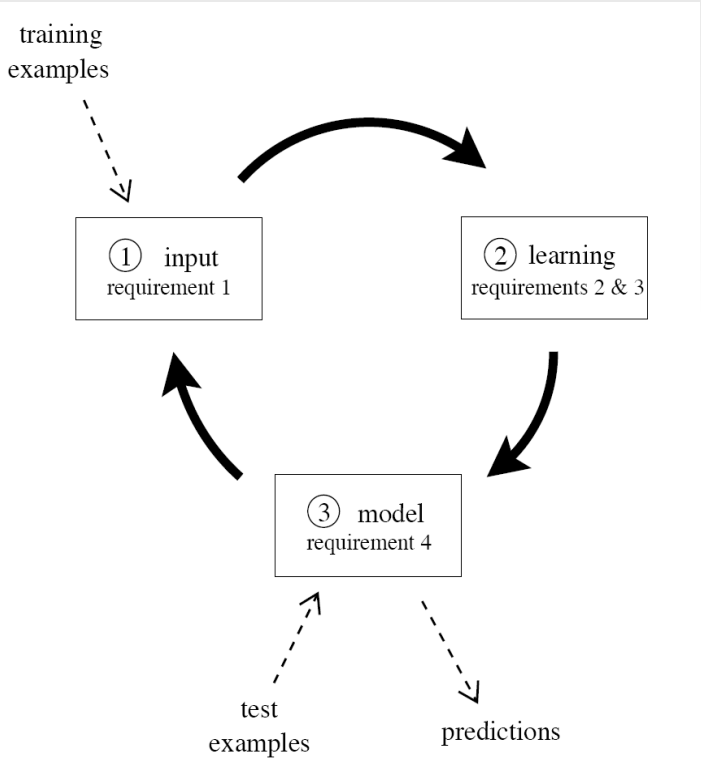
- Sensor Network: 100Kb
- Handheld Computer: 32 Mb
- Server: 400 Mb



Experimental setting

Data Sources

- Random Tree Generator
- Random RBF Generator
- LED Generator
- Waveform Generator
- Function Generator



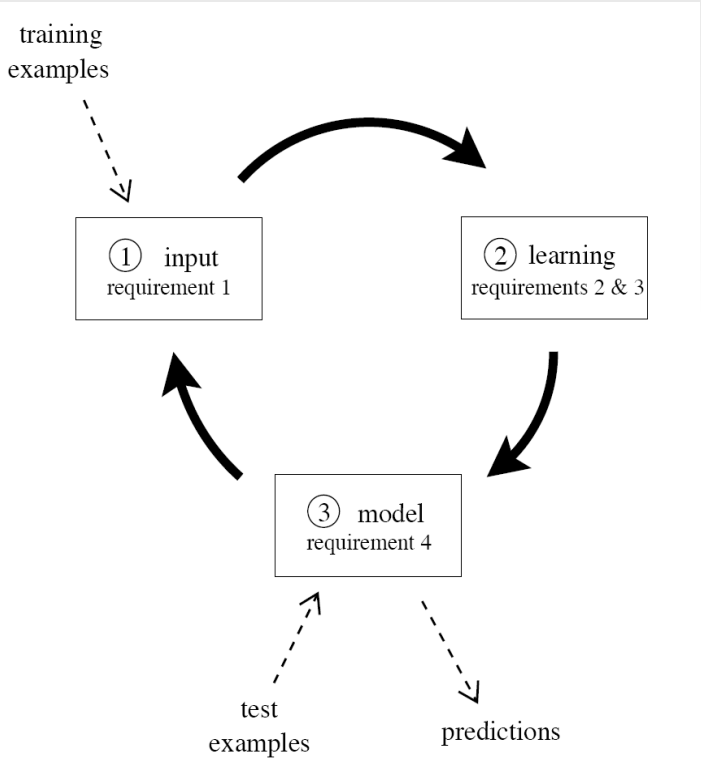
Experimental setting

Classifiers

- Naive Bayes
- Decision stumps
- Hoeffding Tree
- Hoeffding Option Tree
- Bagging and Boosting

Prediction strategies

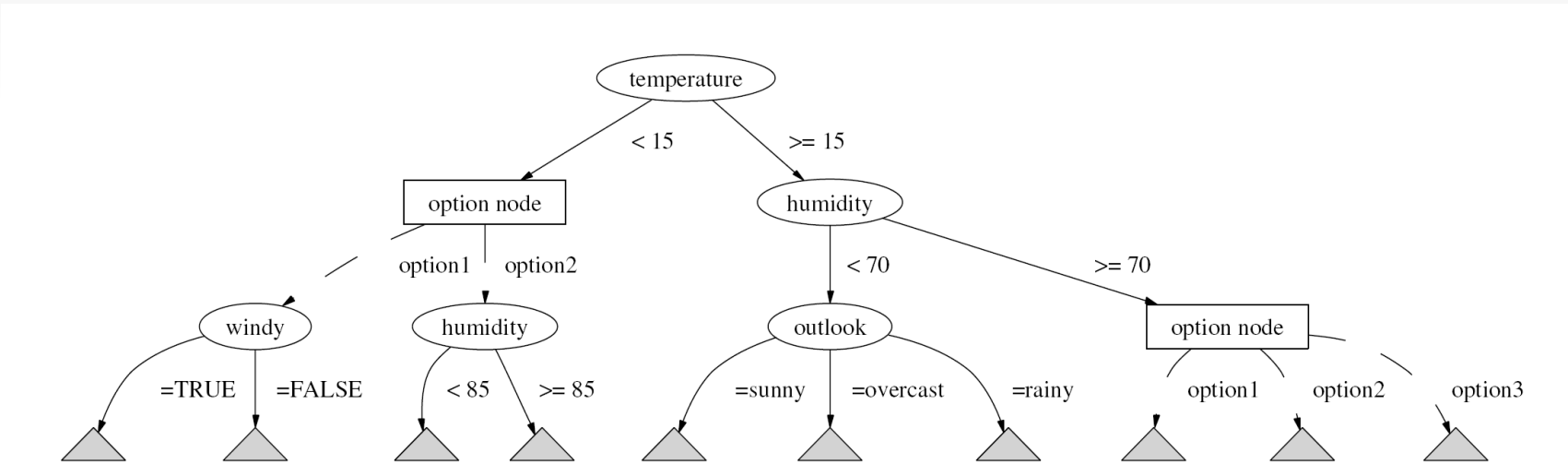
- Majority class
- Naive Bayes Leaves
- Adaptive Hybrid



Hoeffding Option Tree

Hoeffding Option Trees

Regular Hoeffding tree containing additional option nodes that allow several tests to be applied, leading to multiple Hoeffding trees as separate paths.



Extension to Evolving Data Streams

New Evolving Data Stream Extensions

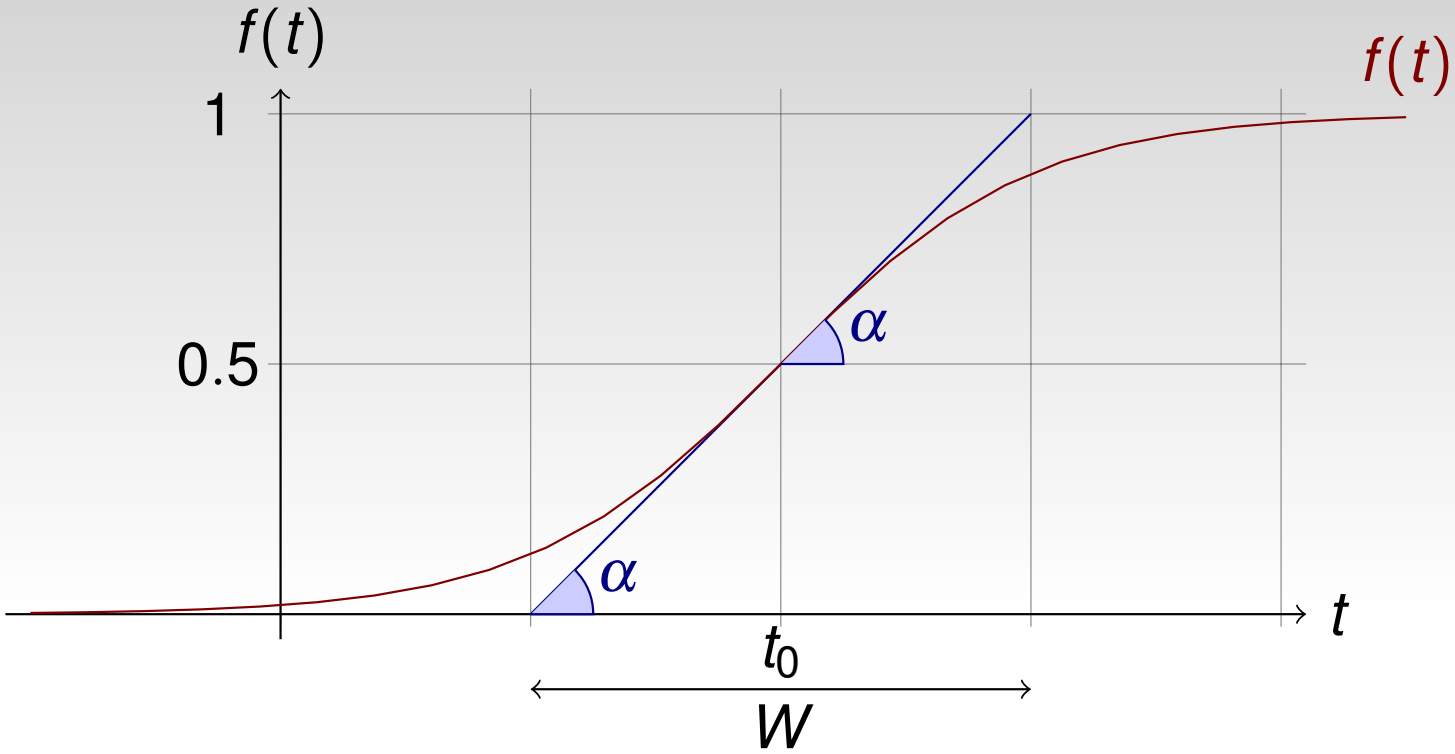
- New Stream Generators
- New UNION of Streams
- New Classifiers

Extension to Evolving Data Streams

New Evolving Data Stream Generators

- Random RBF with Drift
- LED with Drift
- Waveform with Drift
- Hyperplane
- SEA Generator
- STAGGER Generator

Concept Drift Framework

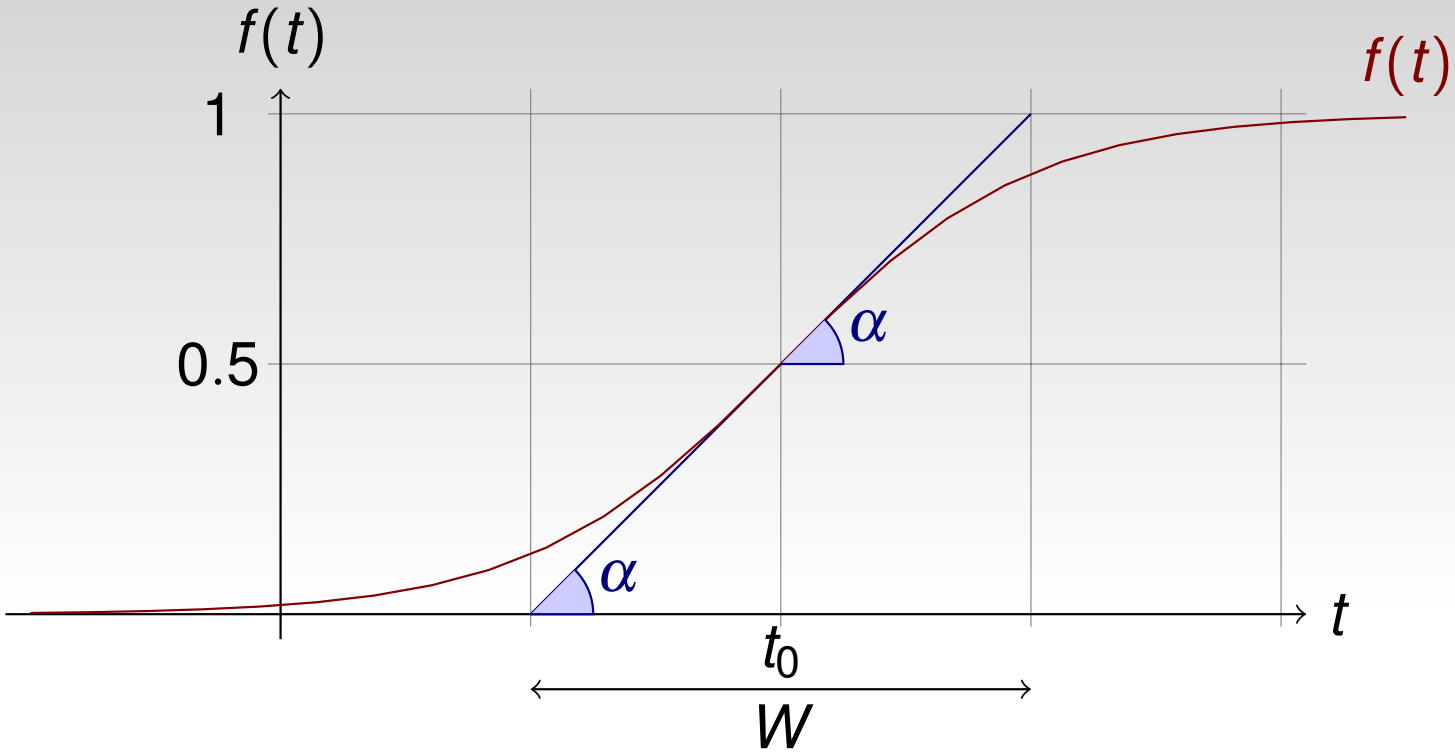


Definition

Given two data streams a, b, we define $c = a \oplus_{t_0}^{W} b$ as the data stream built joining the two data streams a and b

- $\Pr[c(t) = b(t)] = 1/(1 + e^{-4(t-t_0)/W})$.
- $\Pr[c(t) = a(t)] = 1 - \Pr[c(t) = b(t)]$

Concept Drift Framework



Example

- $(((a \oplus_{t_0}^{W_0} b) \oplus_{t_1}^{W_1} c) \oplus_{t_2}^{W_2} d) \ldots$
- $(((SEA_9 \oplus_{t_0}^{W} SEA_8) \oplus_{2t_0}^{W} SEA_7) \oplus_{3t_0}^{W} SEA_{9.5})$
- $\text{CovPokElec} = (\text{CoverType} \oplus_{581,012}^{5,000} \text{Poker}) \oplus_{1,000,000}^{5,000} \text{ELEC2}$

Extension to Evolving Data Streams

New Evolving Data Stream Classifiers

- Adaptive Hoeffding Option Tree
- DDM Hoeffding Tree
- EDDM Hoeffding Tree
- OCBoost
- FLBoost

Ensemble Methods

New ensemble methods:

- Adaptive-Size Hoeffding Tree bagging:
  - each tree has a maximum size
  - after one node splits, it deletes some nodes to reduce its size if the size of the tree is higher than the maximum value
- `ADWIN` bagging:
  - When a change is detected, the worst classifier is removed and a new classifier is added.

ADWIN Bagging

ADWIN

An adaptive sliding window whose size is recomputed online according to the rate of change observed.

ADWIN has rigorous guarantees (theorems)

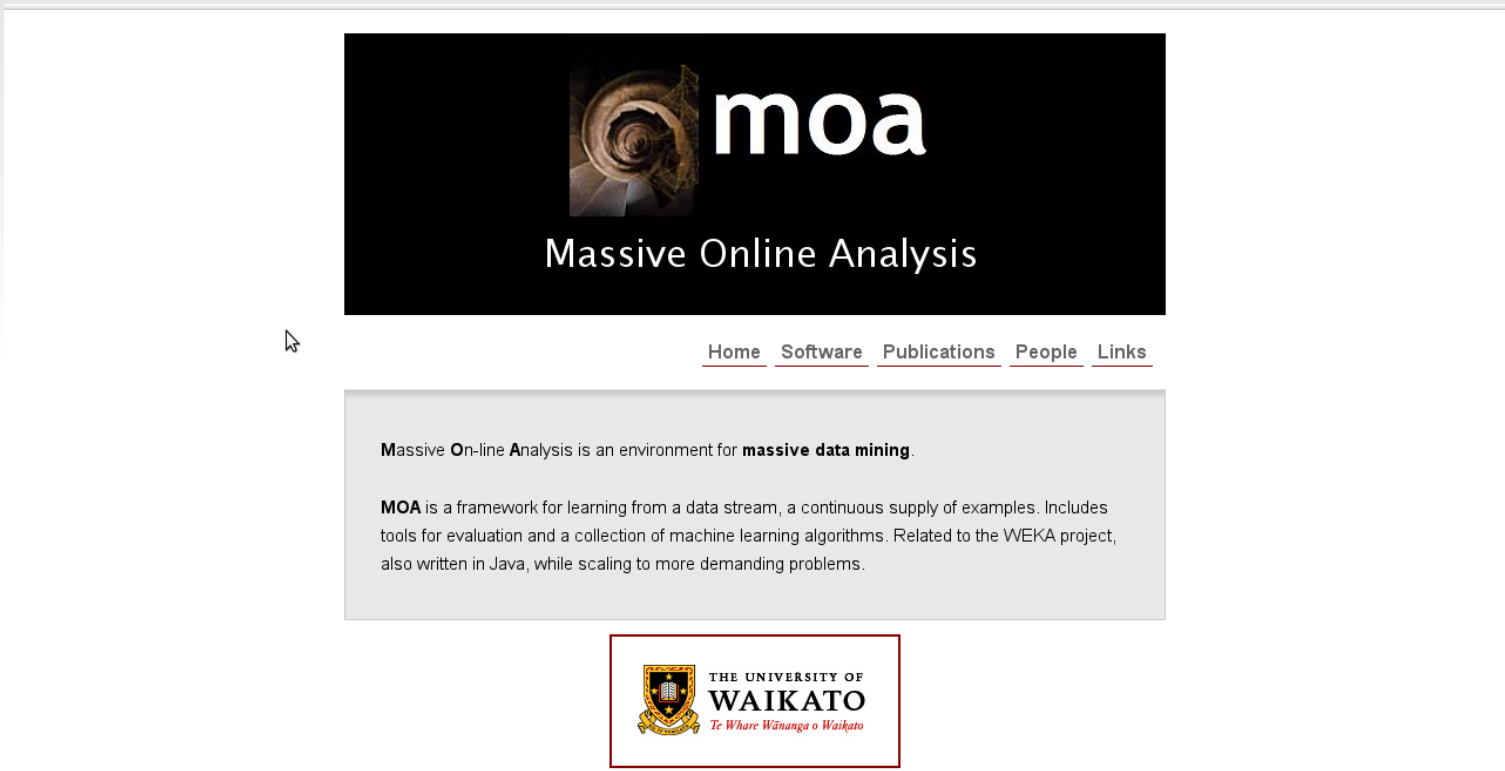
- On ratio of false positives and negatives
- On the relation of the size of the current window and change rates

ADWIN Bagging

When a change is detected, the worst classifier is removed and a new classifier is added.

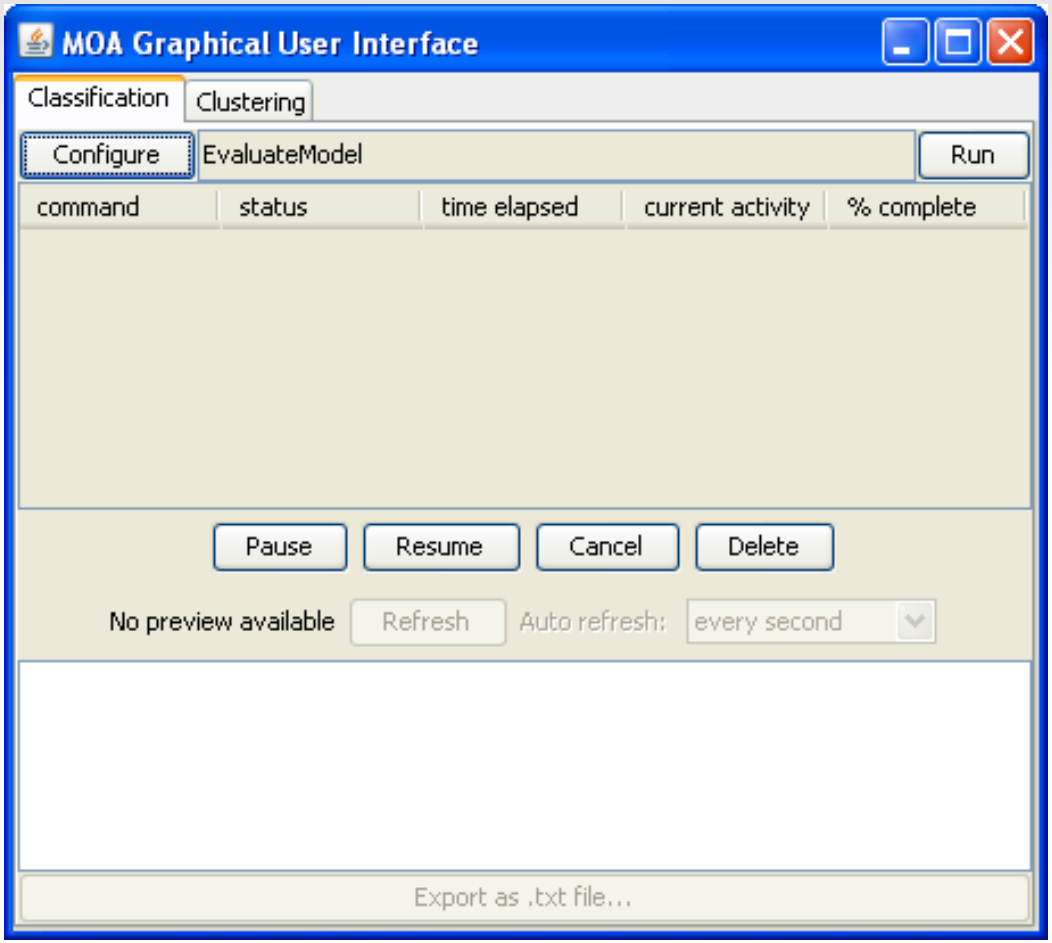
Web

http://www.cs.waikato.ac.nz/~abifet/MOA/



GUI

```
java -cp .:moa.jar:weka.jar
-javaagent:sizeofag.jar moa.gui.GUI
```



Command Line

EvaluatePeriodicHeldOutTest

```
java -cp .:moa.jar:weka.jar -javaagent:sizeofag.jar
moa.DoTask "EvaluatePeriodicHeldOutTest
-l DecisionStump -s generators.WaveformGenerator
-n 100000 -i 100000000 -f 1000000" > dsresult.csv
```

This command creates a comma separated values file:

- training the DecisionStump classifier on the WaveformGenerator data,
- using the first 100 thousand examples for testing,
- training on a total of 100 million examples, and
- testing every one million examples:

Easy Design of a MOA classifier

- `void resetLearningImpl ()`
- `void trainOnInstanceImpl (Instance inst)`
- `double[] getVotesForInstance (Instance i)`
- `void getModelDescription (StringBuilder out, int indent)`

Example: Sentiment Analysis on Twitter

Sentiment analysis

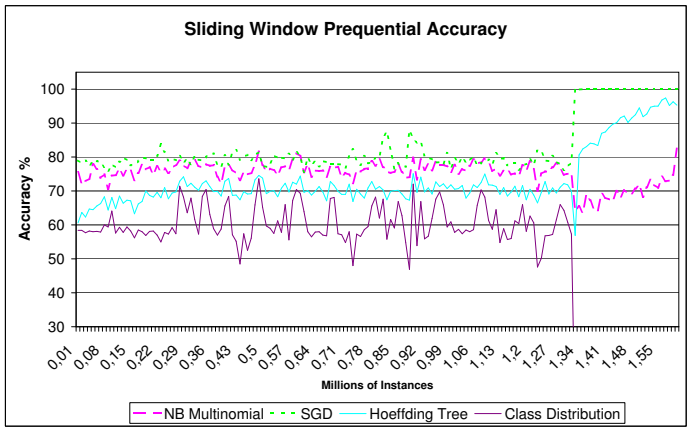
Classification problem into two categories depending on whether they convey positive or negative feelings

Emoticons are visual cues associated with emotional states

Positive Emoticons	Negative Emoticons
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Table: List of positive and negative emoticons.

Twitter Empirical evaluation



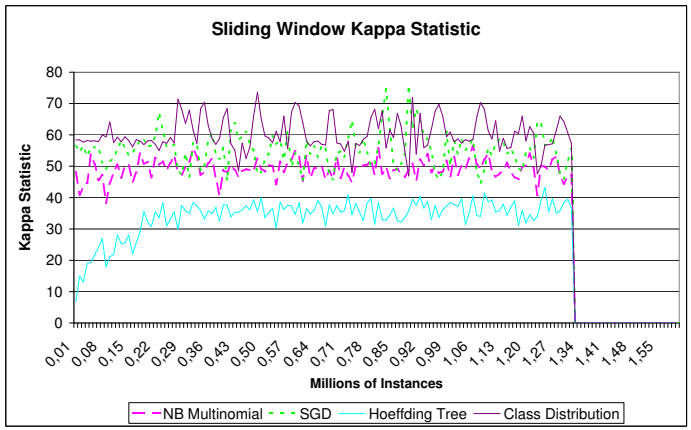
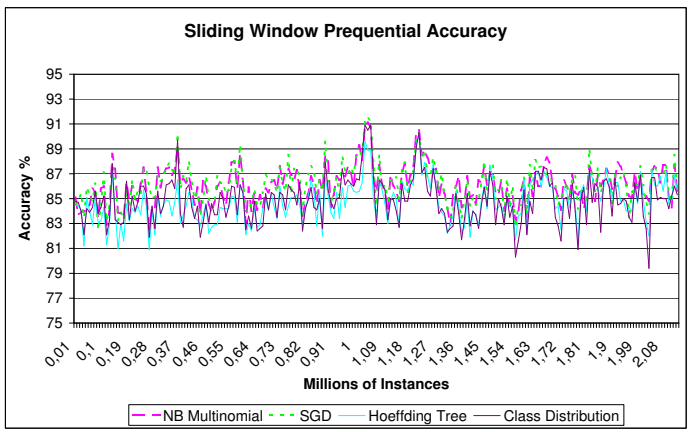


Figure: Accuracy and Kappa Statistic on twittersentiment corpus

Twitter Empirical evaluation



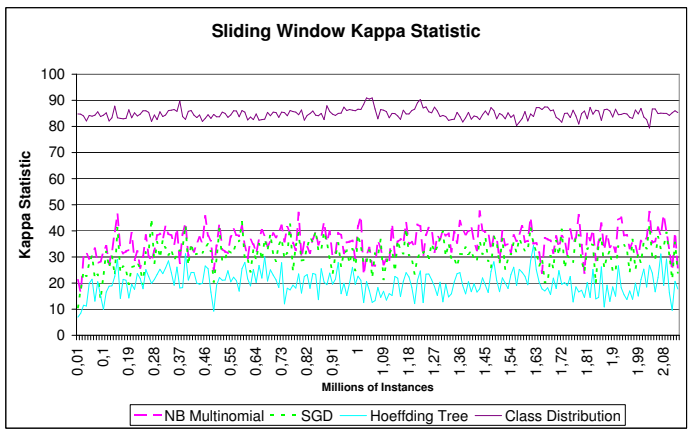


Figure: Accuracy and Kappa Statistic on Edinburgh corpus

Summary

{M}assive {O}nline {A}nalysis is a framework for online learning from data streams.

http://www.cs.waikato.ac.nz/~abifet/MOA/

- It is closely related to WEKA
- It includes a collection of offline and online as well as tools for evaluation:
  - boosting and bagging
  - Hoeffding Trees
- MOA deals with evolving data streams
- MOA is easy to use and extend