MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering

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

University of Waikato
Hamilton, New Zealand

Data Management and Data Exploration Group
RWTH Aachen University, Germany

Cumberland Lodge, 2 September 2010
Workshop on Applications of Pattern Analysis 2010

Mining Massive Data

2007

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

Eric Schmidt, August 2010

Every two days now we create as much information as we did from the dawn of civilization up until 2003.

5 exabytes of data

Twitter

- 106 million registered users
- 3 billion requests a day via its API.

Efficient Algorithms

Evolving Data Streams

Extract information from

- potentially infinite sequence of data
- possibly varying over time
- using few resources

Stream Mining Algorithms

- Fast methods without storing all dataset in memory
- Traditional methods don’t deal with restrictions

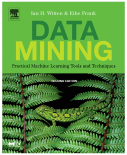
What is MOA?

{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 as well as tools for evaluation:
  - classification
  - clustering
- Easy to extend
- Easy to design and run experiments

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

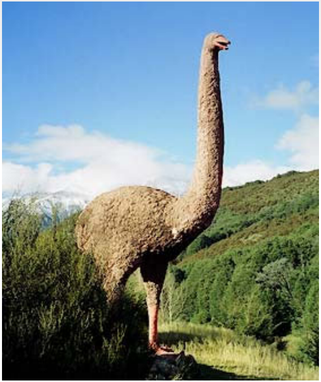

Ian H. Witten & Eibe Frank
DATA
MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION

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

WEKA: the bird

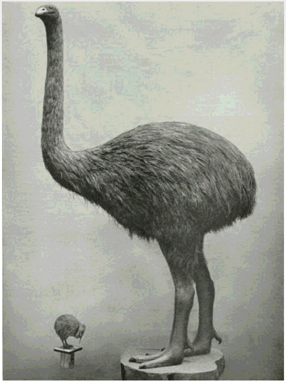
MOA: the bird

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.

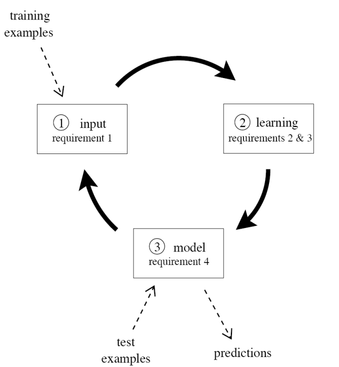


MOA: the bird

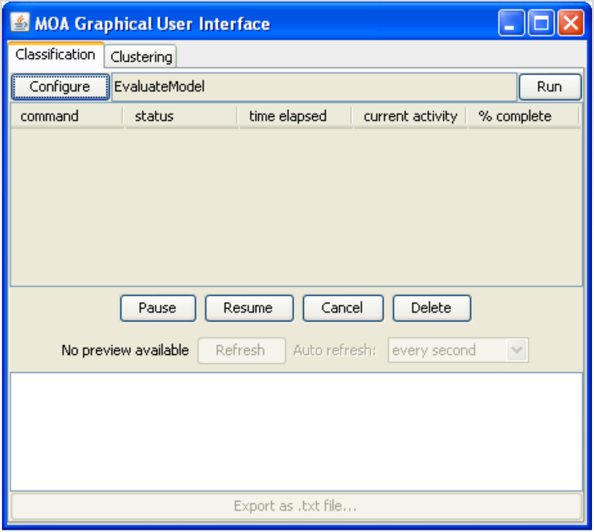
The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

Data stream learning 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



Classification Experimental setting



EvaluateModel

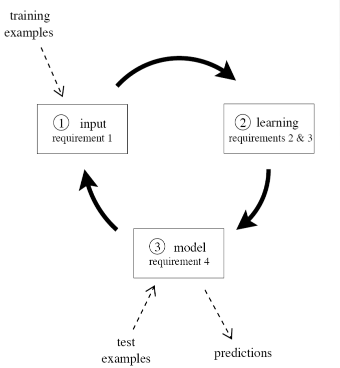
Classification Experimental setting

Evaluation procedures for Data Streams

- Holdout
- Interleaved Test-Then-Train or Prequential

Environments

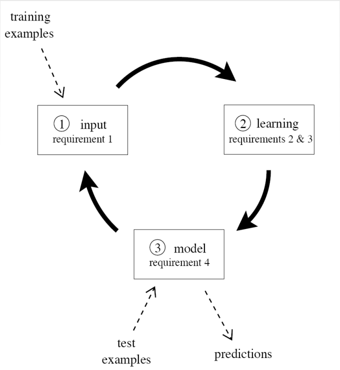
- Sensor Network: 100Kb
- Handheld Computer: 32 Mb
- Server: 400 Mb



Classification Experimental setting

Data Sources

- Random Tree Generator
- Random RBF Generator
- LED Generator
- Waveform Generator
- Hyperplane
- SEA Generator
- STAGGER Generator



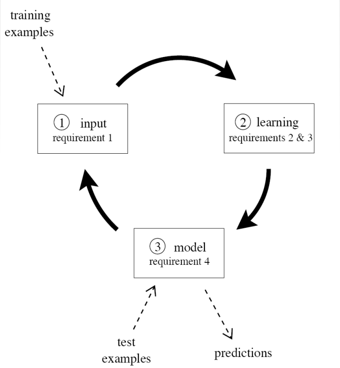
Classification Experimental setting

Classifiers

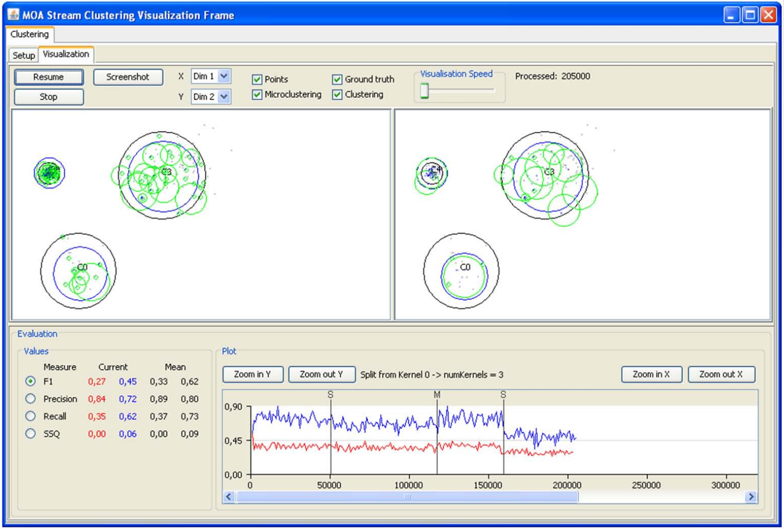
- Naive Bayes
- Decision stumps
- Hoeffding Tree
- Hoeffding Option Tree
- Bagging and Boosting
- ADWIN Bagging and Leveraging Bagging

Prediction strategies

- Majority class
- Naive Bayes Leaves
- Adaptive Hybrid



Clustering Experimental setting



X Dim 1

Y Dim 2

☑ Points ☑ Ground truth

☑ Microclustering ☑ Clustering

Processed: 205000

Measure	Current		Mean	
● F1	0,27	0,45	0,33	0,62
○ Precision	0,84	0,72	0,89	0,80
○ Recall	0,35	0,62	0,37	0,73
○ SSQ	0,00	0,06	0,00	0,09

Split from Kernel 0 -> numKernels = 3

Clustering Experimental setting

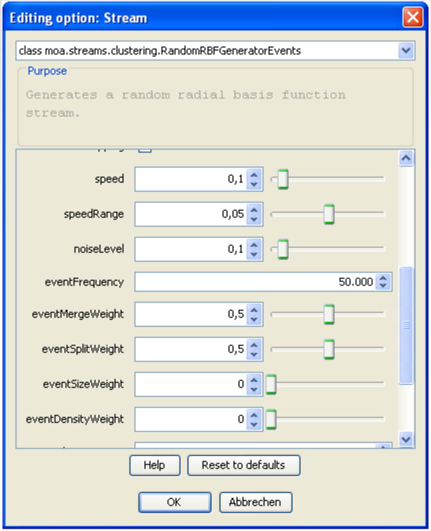
Internal measures	External measures
Gamma	Rand statistic
C Index	Jaccard coefficient
Point-Biserial	Folkes and Mallow Index
Log Likelihood	Hubert Γ statistics
Dunn's Index	Minkowski score
Tau	Purity
Tau A	van Dongen criterion
Tau C	V-measure
Somer's Gamma	Completeness
Ratio of Repetition	Homogeneity
Modified Ratio of Repetition	Variation of information
Adjusted Ratio of Clustering	Mutual information
Fagan's Index	Class-based entropy
Deviation Index	Cluster-based entropy
Z-Score Index	Precision
D Index	Recall
Silhouette coefficient	F-measure

Table: Internal and external clustering evaluation measures.

Clustering Experimental setting

Clusterers

- StreamKM++
- CluStream
- ClusTree
- Den-Stream
- D-Stream
- CobWeb



Editing option: Stream

class moa.streams.clustering.RandomRBFGeneratorEvents

Purpose

Generates a random radial basis function stream.

speed 0,1

speedRange 0,05

noiseLevel 0,1

eventFrequency 50.000

eventMergeWeight 0,5

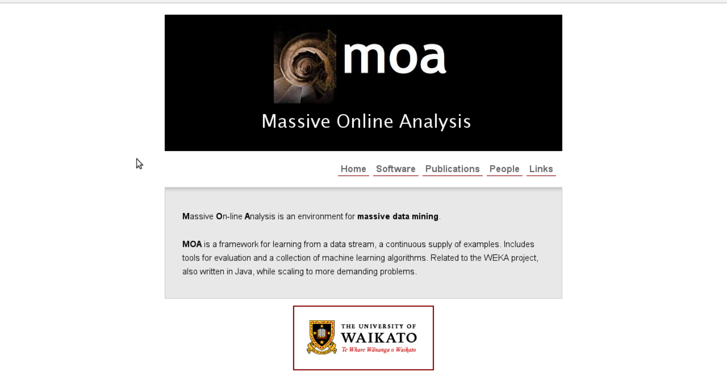
eventSplitWeight 0,5

eventSizeWeight 0

eventDensityWeight 0

Web

`http://www.moa.cs.waikato.ac.nz`

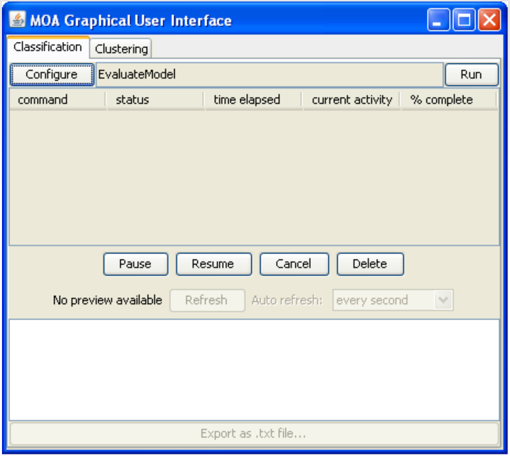


Massive **O**n-line **A**nalysis is an environment for **massive data mining**.

MOA is a framework for learning from a data stream, a continuous supply of examples. Includes tools for evaluation and a collection of machine learning algorithms. Related to the WEKA project, also written in Java, while scaling to more demanding problems.

GUI

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



EvaluateModel

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)`

Easy Design of a MOA clusterer

- `void resetLearningImpl ()`
- `void trainOnInstanceImpl (Instance inst)`
- `Clustering getClusteringResult()`

Extensions of MOA

- Multi-label Classification
- Itemset Pattern Mining
- Sequence Pattern Mining

Summary

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

`http://www.moa.cs.waikato.ac.nz`

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

SCORE.
POINTS
N.Z. 829.
Eng. 39.
N.Z.
THE MOA AND THE LION.