Accuracy Updated Ensemble for Data Streams with Concept Drift

Dariusz Brzeziński and Jerzy Stefanowski

Poznan University of Technology

Outline

- Data streams
- Concept drift
- Accuracy Updated Ensemble
- Experimental evaluation
- MOA
- Future work

Data streams

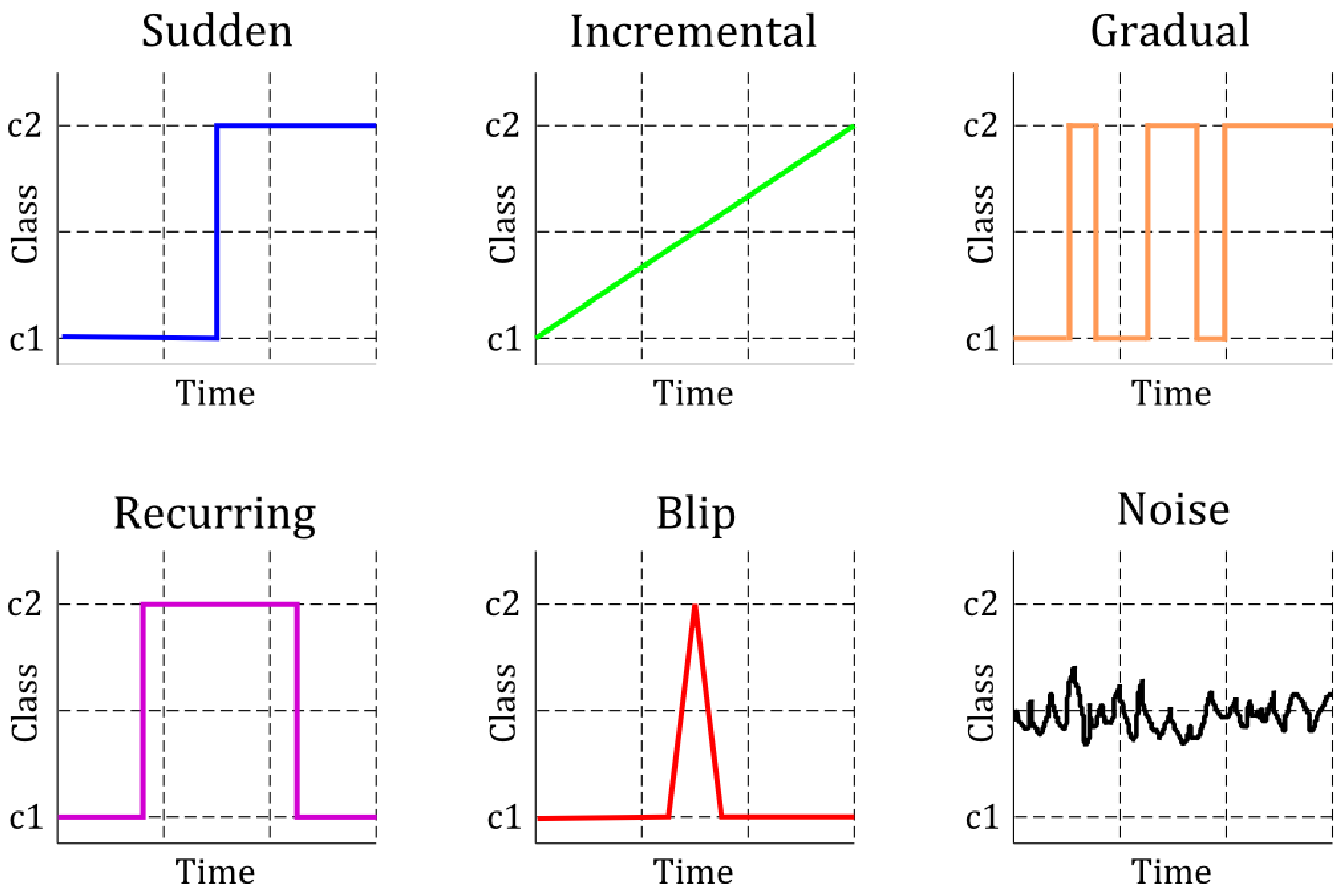
- The ”Digital Universe” in 2007 was estimated to be 281 exabytes large
- The amount of data created exceeds available storage
- Incoming tuples processed as a stream of data

New challenges for data mining algorithms!

Data stream contraints

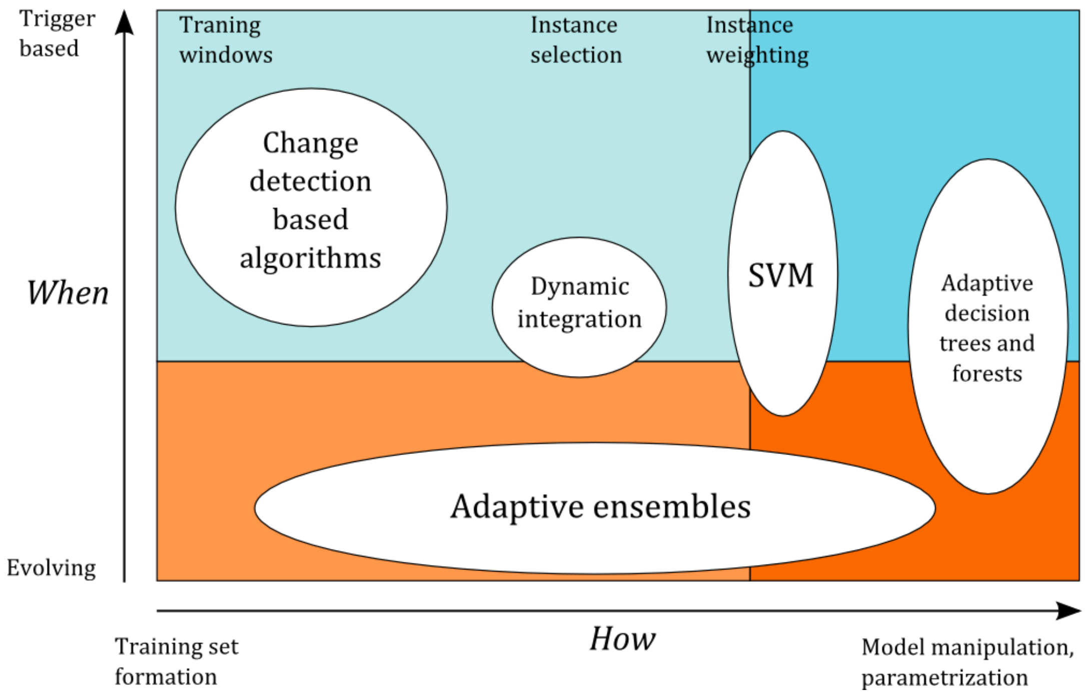
- Limited time
  - examples arrive rapidly
  - each example can be processed only once
- Limited memory
  - streams are too large to be processed as a whole
- Concept drift
  - data streams can evolve over time
  - changes that are unpredictable (not seasonal) are called *Concept drift*

Types of concept drift



Stream data mining algorithms

- Drift detectors
- Forgetting mechanisms



Single classifiers: DDM, EDDM, VFDT, FISH, FLORA, ADWIN
Ensemble classifiers: SEA, AWE, HOT, Online Bagging, ASHT

Accuracy Weighted Ensemble

„Mining concept-drifting data streams using ensemble classifiers ”, H. Wang et al.; KDD 2003

Idea:

Weight classifiers according to the current data distribution

- Formal proof that classifiers weighted this way are equally or more accurate than classifiers built upon all examples without weights
- Weights approximated by computing classification error on the most recent data chunk

AWE drawbacks

- Accuracy is highly dependent on chunk size
- Poorer accuracy for data streams with slow gradual concept drift
- Sudden concept drifts can sometimes mute all base classifiers

Accuracy Updated Ensemble

Idea:

Incrementally update base classifiers according to the current distribution while keeping them diversified.

- Inspired by AWE’s weighting mechanism
- Chunk size independent
- More accurate
- Reacts better to concept drift

Accuracy Updated Ensemble

- AWE inspired:
  - using mean square error on the most recent data chunk to weight component classifiers
- New elements:
  - Hoeffding Trees as base classifiers
  - updating component classifiers according to their weight
  - diversifying components
  - preventing classifier muting $(w_i = \frac{1}{MSE_i + \varepsilon})$

Experiments

- 4 algorithms: HT+Win, HOT, AWE, AUE
- 3 real and 4 artificial data sets
- From 2.5 thousand do 10 million examples
- Gradual and sudden concept drift
- Classifiers were evaluated using chunks of data:
  - Test and train time
  - Memory usage
  - Accuracy

Results

Table: Results for Donation data set

	Chunk Training	Chunk Testing	Accuracy	Memory
HOT	5,17 s	0,01 s	85,07%	18,49 MB
AWE	0,04 s	0,01 s	70,38%	0,17 MB
HT+Win	0,02 s	0,01 s	79,08%	0,18 MB
AUE	0,24 s	0,05 s	84,72%	0,86 MB

Results

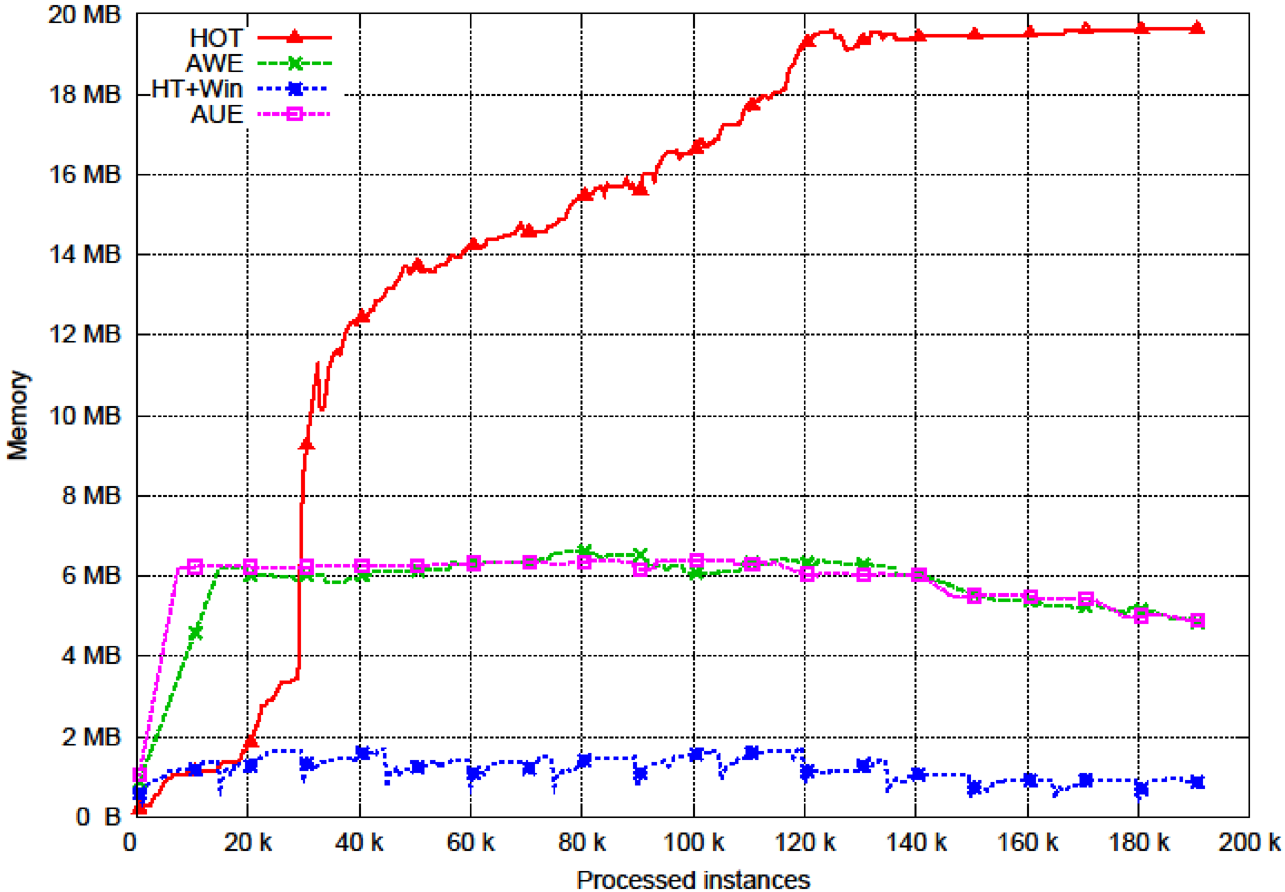


Figure: Memory usage on the Donation data set

Results

- Sliding window, AWE:
  - least accurate
  - least resource consuming
- HOT:
  - time and memory requirements grew linearly with each data chunk
- AUE:
  - as accurate as HOT
  - constant time and memory
  - much more accurate than AWE

{M}assive {O}nline {A}nalysis

- Framework for online learning from data streams
- Closely related to WEKA
- Contains:
  - classifiers
  - clustering algorithms
  - stream generators
- Easy to extend
- AWE, AUE, and Data Chunk Evaluation are included in the latest release

Summary

- A comparison of chunk ensemble methods
- AUE a new classifier:
  - constant time and memory
  - reacts to concept drift
  - as accurate as more expensive methods
- Three algorithms contributed to MOA
- Plans to add more diversity and a pruning mechanism

Thank you!