Economically-Efficient Data Stream Analysis

Roberto Oliveira Jr.
Advisor: Adriano Veloso
Co-advisor: Wagner Meira Jr.

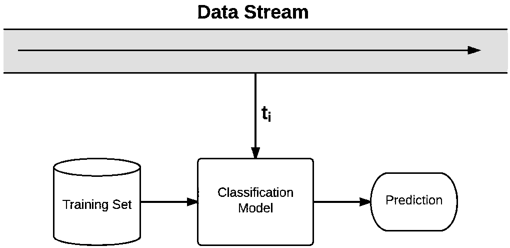
Computer Science Dept - UFMG - Brazil

Data Stream

- Definition
  - Fast and possible unbounded sequence of data that arrives at time-varying.
- Motivation
  - It allows us to process huge volumes of data.
- Problem
  - Automatically extraction of relevant patterns and relations from data that is continuously created.

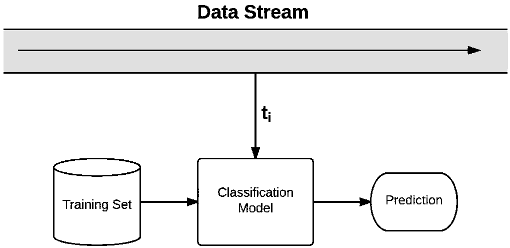
Classification in Data Streams

- Classification models are applied to distinguish between pre-defined labels.



Classification in Data Streams

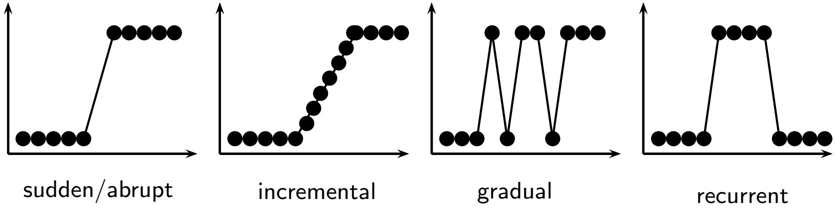
- Classification models are applied to distinguish between pre-defined labels.



- Data characteristics may change with time.

Concept Drifts

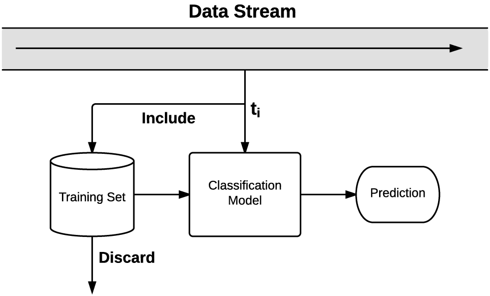
- Concept Drift is unforeseen changes in data's nature over time.



- Data streams contains combination of such patterns.

Classifying Data Streams

- Effective classification requires:
    - Updating the classification model as the stream evolves.
        - Taking into account resources limitation: memory, time and learning requirements.



Research Question

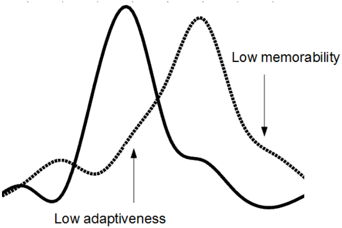
How to deal with concept drifts?

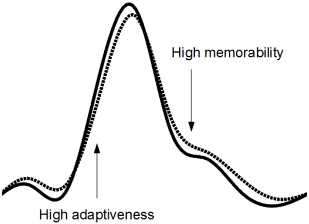
Classification Model

- Classification models are composed by association rules.
  - $\{x \to y\}$, where $x \in X$ and $y \in Y$
- Efficiently updated as the training set evolves.
- Models are built on-the-fly:
  - For a given $[x_i, *]$, rules $\{x \to y\}$ such that $x \in x_i$ are produced.
  - Prediction is performed from the combination of these rules.
- At each time step is produced a model $\mathcal{R}(x_i)$.

Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
  - Adaptiveness:
    - The ability to adapt itself to drifts.
  - Memorability:
    - The ability to recover itself from drifts.



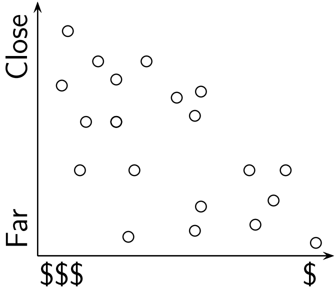


Dealing with Drifts

- Improving both properties simultaneously may lead to a conflict-objective problem.
    - Improve adaptiveness may hurt memorability, and vice-versa.

Economic Efficiency

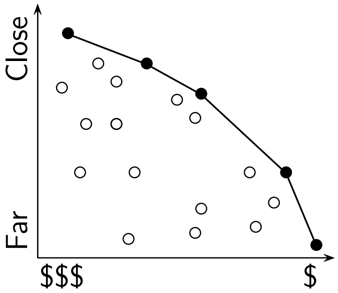
Example: Hotels in Petrópolis.



Pareto Efficiency

Pareto frontier: A point a is said to dominate b iff both of the following conditions are hold:

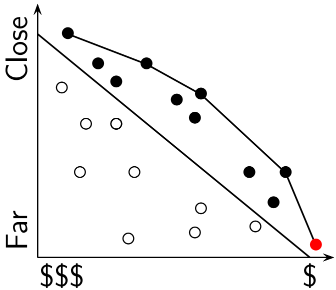
- $U_c(a) \geq U_c(b)$ and $U_d(a) \geq U_d(b)$
- $U_c(a) > U_c(b)$ or $U_d(a) > U_d(b)$



Compensation — Kaldor-Hicks Principle

Region of compensation:

- Overall utility: $U(d_i) = U_m(d_i) + U_a(d_i)$
- Baseline point:
  $d^* = \{d_i \in \mathcal{P}_n | \forall d_j \in \mathcal{P}_n : U(d_i) \leq U(d_j)\}$

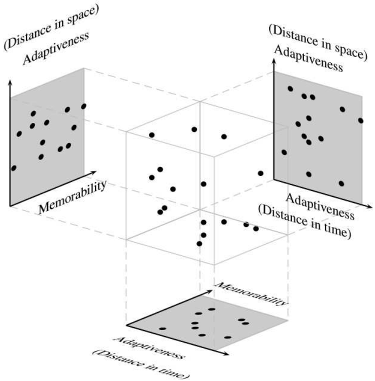


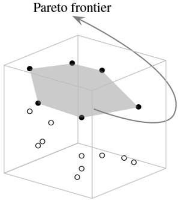
Utility Measures

- Distance in space:
  - How similar training instance t_j is to the newest instance t_n.
  - $U_s(t_j) = \frac{|\mathcal{R}(t_n) \cap \mathcal{R}(t_j)|}{|\mathcal{R}(t_n)|}$
- Distance in time:
  - How fresh is the training instance.
  - $U_t(t_j) = \frac{\gamma(t_j)}{\gamma(t_n)}$.
    - $\gamma(t_j)$ returns the time in which training instance t_j arrived.
- Random permutation of training instances:
  - $U_r(t_j) = \frac{\alpha(t_j)}{|\mathcal{D}_n|}$
    - $\alpha(t_j)$ returns the position of t_j in the shuffle.
    - $\mathcal{D}_n$ is the training set at time step n.

Utility Space

- At each time step n:
  1. Place training instances in the utility space.
  2. Select training instances in the Efficiency Region:
     - Pareto-Efficient Selective Sampling (PESS).
     - Kaldor-Hicks Selective Sampling (KHSS).

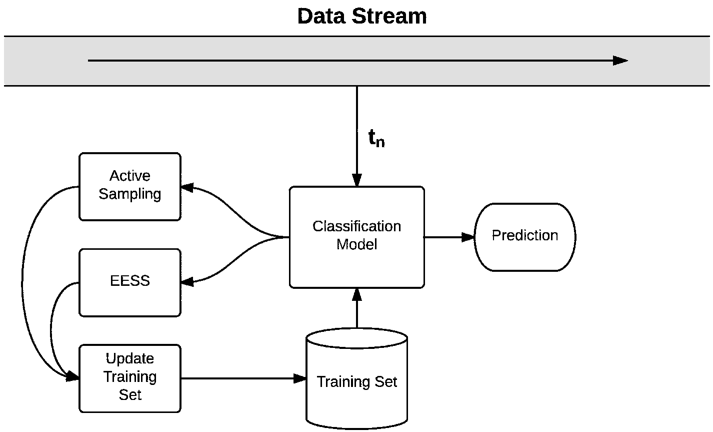




Reducing Labeling Efforts

- Random Active Learning
  - Naive strategy.
  - Simple to integrate.
  - Labeling Effort control: β.

Economically-Efficient Selective Sampling



Experimental Evaluation

Setup

- Interleaved Test-Then-Train
- 1% of data provided as training seed;
- Massive Online Analysis (MOA) framework as evaluation environment;
- Baselines:

Algorithm	Adaptiveness	Memorability
AC (KDD 2011)	Active Learning	Base Learner
HAT (JMLR 2011)	ADWIN	Trees Ensemble
ILAC (SIGIR 2011)	Data Projection	Incremental Training Set

Evaluation

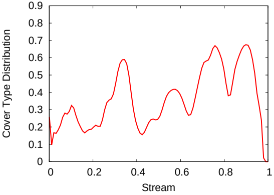
- Measures used:
  - Mean Squared Error.
  - Labeling Effort: 10%; **25%**; 50%; 75% and 100%;
    - AC and EESS.
  - Training set size.
  - RAM-Hours:
    - A GB of RAM deployed for 1 hour execution.
- Datasets:

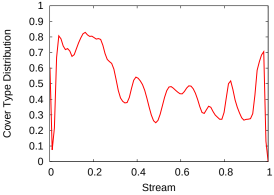
	Concept Drift Pattern			
Dataset	Sudden	Incremental	Gradual	Recurrent
Presidential Elections	-	X	X	-
Person of the Year	-	X	X	-
FIFA World Cup - EN	X	-	-	-
FIFA World Cup - PT	X	-	-	-
Cover Type	X	-	X	X
Spam Filtering	X	-	X	X
Poker Hand	-	-	X	X

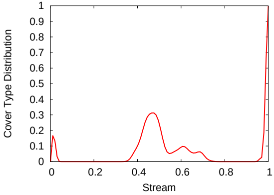
Evaluation

Forest Cover Type Prediction

- Data from forest cover type in United State territory;
- 581,102 examples with 54 features distributed among 7 classes;
- Concept Drifts: Sudden, Gradual, Recurrent;



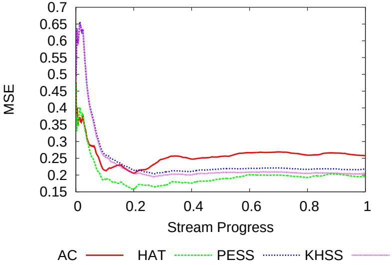


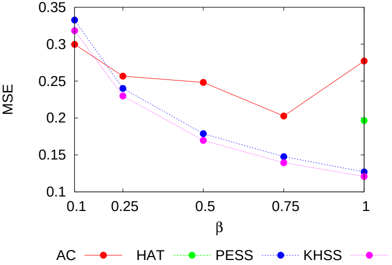


Evaluation

Forest Cover Type Prediction

MSE and Labeling Efforts

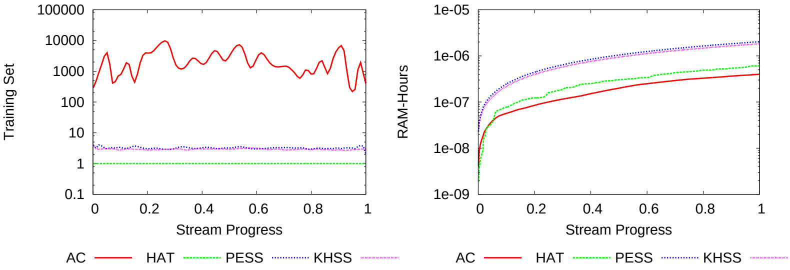




Evaluation

Forest Cover Type Prediction

Training Size and RAM-Hours



Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.

Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.
- Efficiency and accuracy.
  - Incremental classifiers.
  - Adaptiveness and Memorability.
  - Pareto efficiency and compensation principle.
  - Simple-to-compute utility measures.
  - Ours algorithms shown to be robust in different scenarios.

Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.
- Efficiency and accuracy.
  - Incremental classifiers.
  - Adaptiveness and Memorability.
  - Pareto efficiency and compensation principle.
  - Simple-to-compute utility measures.
  - Ours algorithms shown to be robust in different scenarios.
- Future work includes:
  - Other utility measures.
  - Employ our method to reduce Labeling Efforts.
  - Explore other classification models.

Thank you!

robertolojr@dcc.ufmg.br