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.
    - Keep track of data streams is useful for systems monitoring, online social network advertising, etc.

Social Networks Streams and Advertising

Superbowl 2013

Sentiment Streams and Advertising

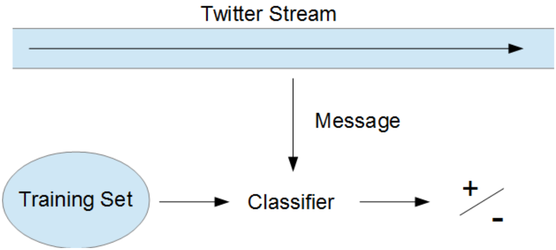
and last year ... Brazil 1 $\times$ Germany 7

PlayStation Brasil @PlayStation_BR · 2 h
#SeFosseNoPLAY era apertar o Reset e começar outra!
#BRA🇧🇷 vs #GER🇩🇪 pic.twitter.com/wMtwpsGNFF

Just hit the reset button and start again
#BRA🇧🇷 vs #GER🇩🇪

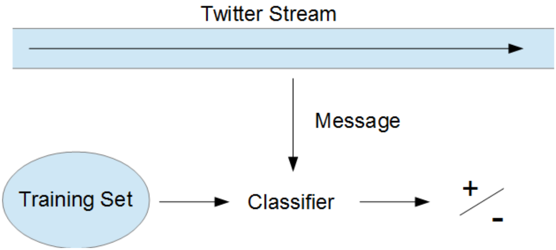
Classifying Data Streams

- Classifiers are applied to distinguish between pre-defined labels.



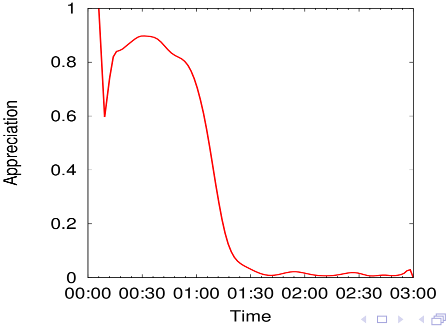
Classifying Data Streams

- Classifiers are applied to distinguish between pre-defined labels.

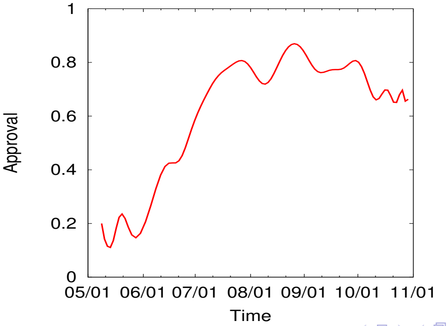


- Data characteristics may change with time.

Sports (WC 2010)

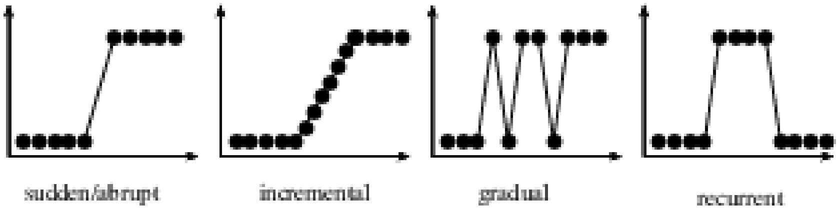


Elections (Brazil 2010)



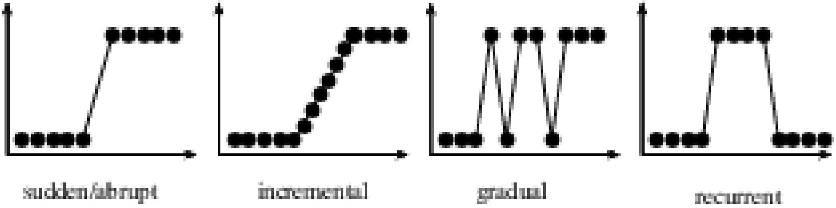
Concept Drifts Types

- Most common types of Concept Drifts:



Concept Drifts Types

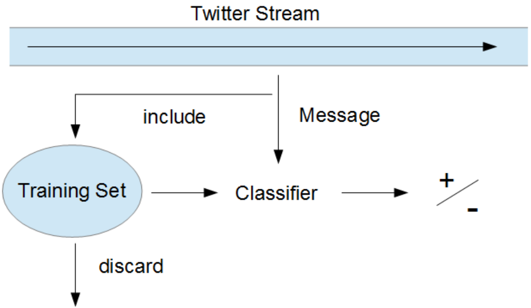
- Most common types of Concept Drifts:



- Data streams contains combination of such concept drifts types.

Classifying Data Streams

- Effective classification requires:
  - Updating the training-set to mitigate drifts.
  - Updating the classifier accordingly.
    - Limited resources: memory, time and learning requirements.



Research Questions

1. Resources:
   - How to build classification models fast?
2. Accuracy:
   - How to deal with concept drifts?
3. Effort:
   - How to reduce labeling effort?

Research Questions

1. Resources:
   - How to build classification models fast?
2. Accuracy:
   - How to deal with concept drifts?
3. Effort:
   - How to reduce labeling effort?

Classification Model

- Our classification model $\mathcal{R}$ is composed of rules $\{X \to c_i\}$.
  - X is any combination of features in the target instance.
  - c_i is a label.

Label Scoring

- Given a target instance t_n:
  1. Build a classification model $\mathcal{R}(t_n)$ composed of rules $\{X \rightarrow c_i\}$ for which $X \subseteq t_n$.
  2. The label score is given by the linear combination of the rules.

Label Scoring

- Given a target instance t_n:
  1. Build a classification model $\mathcal{R}(t_n)$ composed of rules $\{X \rightarrow c_i\}$ for which $X \subseteq t_n$.
  2. The label score is given by the linear combination of the rules.

"It works well, my sister loves it, but unfortunatelly it broke."

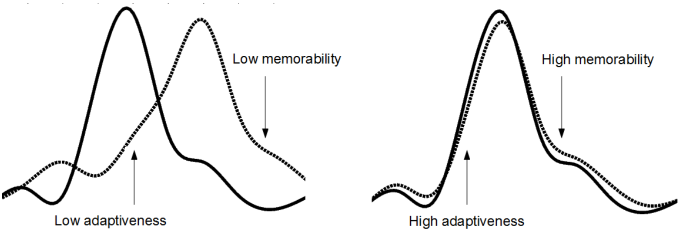
- {broke} → negative (0.77)
- {work, well} → positive (0.85)
- {love, it} → positive (0.91)

Research Questions

1. Resources:
   - How to build classification models fast?
2. Accuracy:
   - How to deal with concept drifts?
3. Effort:
   - How to reduce labeling effort?

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.

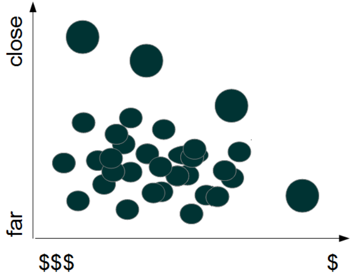


Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
  - Adaptiveness:
    - The ability to adapt itself to drifts.
    - The training-set must contain fresh messages.
  - Memorability:
    - The ability to recover itself from drifts.
    - The training-set must contain pre-drift messages.
- Improving both properties simultaneously may lead to a conflict-objective problem.
  - Improve adaptiveness may hurt memorability, and vice-versa.

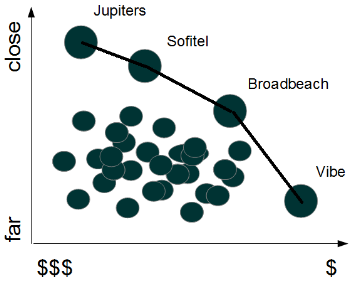
Pareto Efficiency

Example: hotels in Petrópolist.



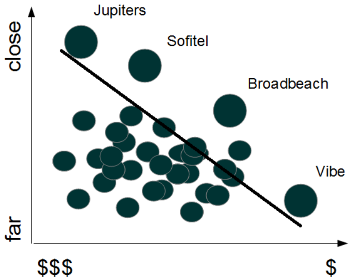
Pareto Efficiency

Pareto frontier.



Compensation — Kaldor-Hicks Principle

Region of compensation.

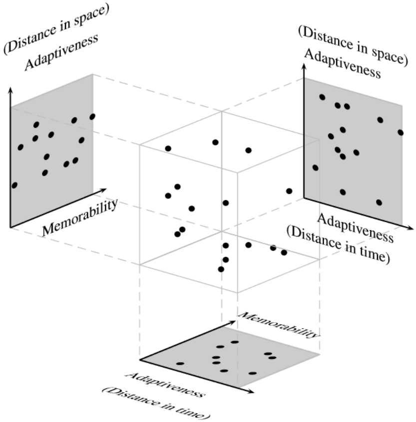


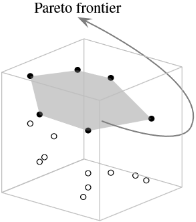
Utility Measures

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

Utility Measures

1. At each time step *n*:
   1. Place candidate messages in the utility space.
   2. Select messages in the Pareto frontier.





Research Questions

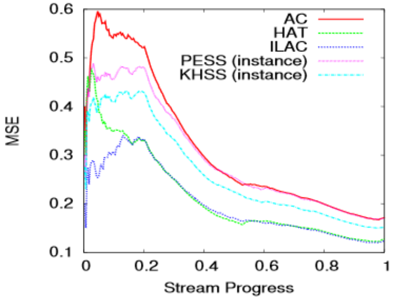
1. Resources:
   - How to build classification models fast?
2. Accuracy:
   - How to deal with concept drifts?
3. Effort:
   - How to reduce labeling effort?

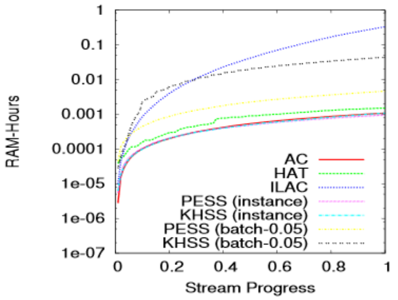
Evaluation

- Measures used:
  - Mean Squared Error.
  - RAM-Hours:
    - A GB of RAM deployed for 1 hour execution.
- Labeling Effort:
  - Different batch sizes and δ values.
- Three datasets:
  - Brazilian elections 2010.
  - World Cup 2010.
  - Person of the Year 2010 (Assange vs. Zuckerberg)
  - Baselines:
    - AC – Active Classifiers (KDD 2011)
    - HAT – Hoeffding Adaptive Trees (JMLR 2011)
    - ILAC – Incremental Lazy Classifiers (SIGIR 2011)

Evaluation

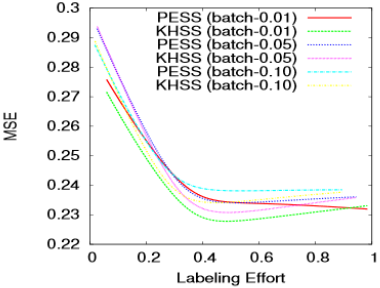
- MSE and RAM-Hours

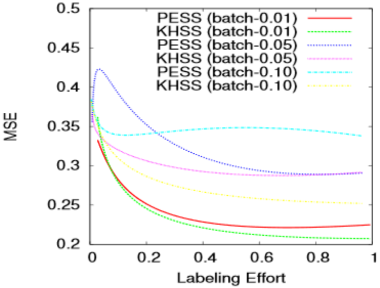




Evaluation

- MSE and Labeling Effort





Conclusions

- Sentiment analysis on Twitter streams.
  - Limited computing and training resources.
  - Sentiment drifts.
- Efficiency and accuracy.
  - Incremental classifiers.
  - Pareto efficiency and compensation principle.
- Our results.
  - 50% reduction in terms of labeling effort without impact on accuracy.
- Future work includes:
  - Other utility measures.
  - Other application scenarios.

Thank you!

adrianov@dcc.ufmg.br