Economically-Efficient Data Stream Analysis

Roberto Oliveira Jr.
Orietador: Adriano Veloso
Co-orientados: Wagner Meira Jr.

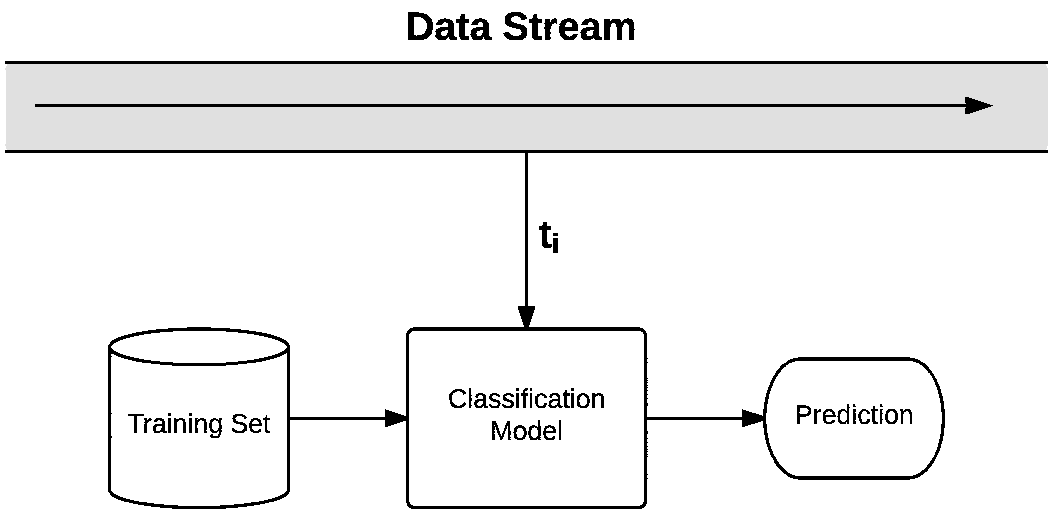
DCC - UFMG - Brazil

Fluxo de Dados

- Definição
  - Sequência de dados possivelmente ilimitada, de alta velocidade onde os dados chegam em intervalos de tempos variados.
- Motivação
  - Permite o processamento de grandes volumes de dados.
- Problema
  - Extrair automaticamente padrões e relações relevantes de dados continuamente criados.

Classificação em Fluxo de Dados

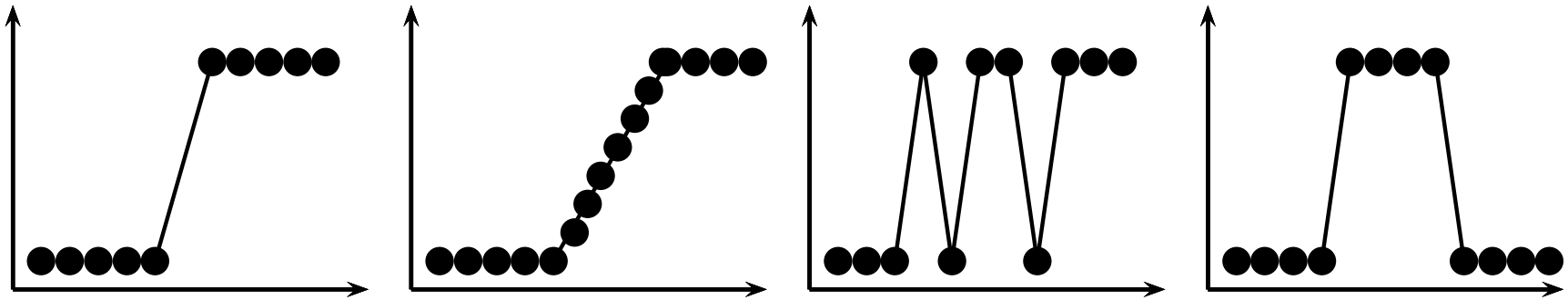
- Modelos de classificação são aplicados para distinguir entre rótulos pré-estabelecidos.



- As características dos dados podem mudar ao longo do tempo.

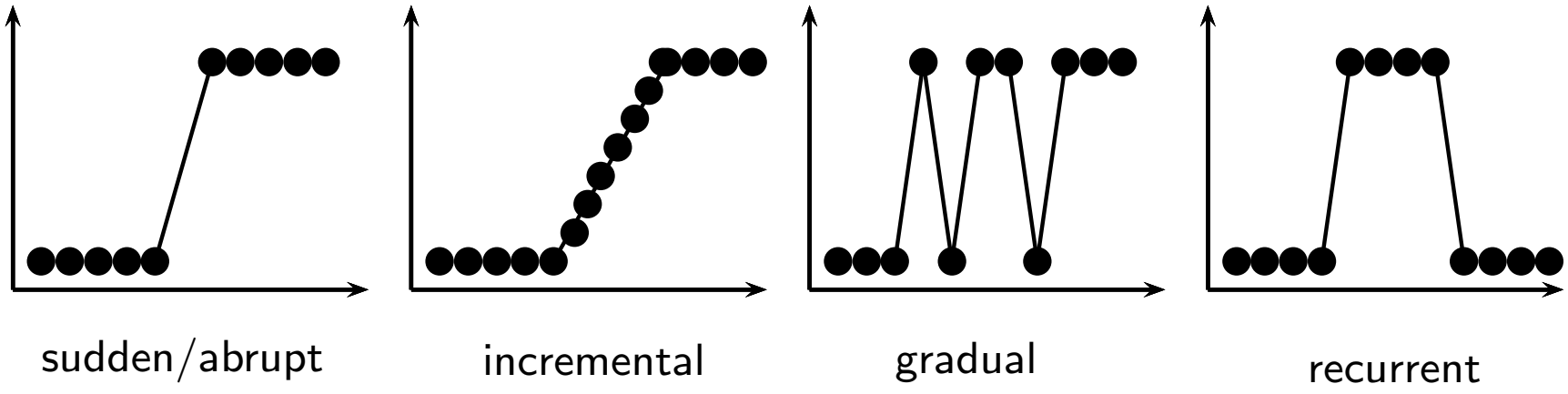
Mudanças de Conceito

- Mudança de Conceito é a alteração imprevisível da natureza dos dados ao longo do tempo.



Mudanças de Conceito

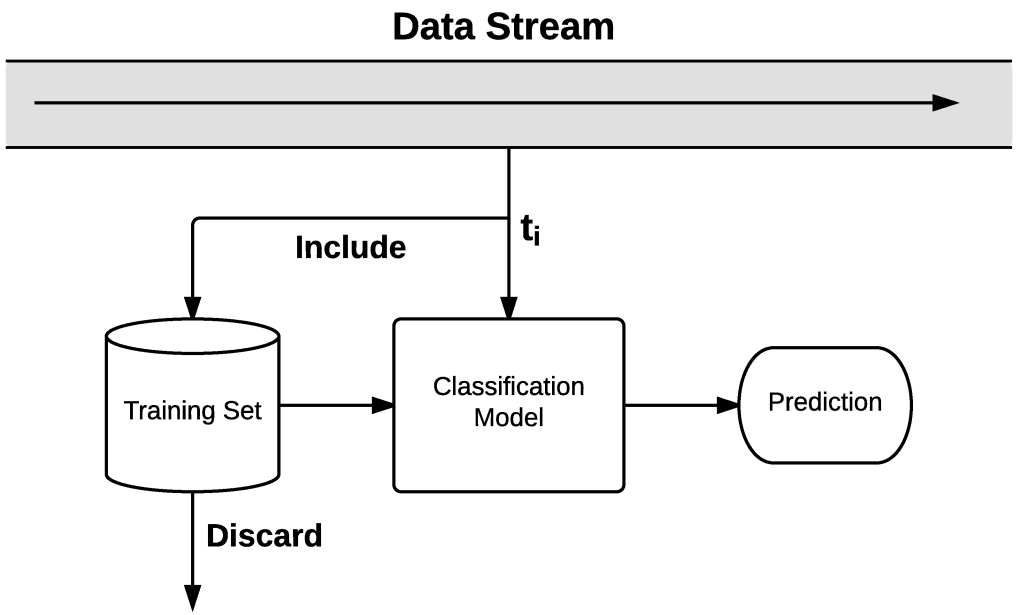
- Mudança de Conceito é a alteração imprevisível da natureza dos dados ao longo do tempo.



- Fluxo de dados contém combinação destes padrões.

Classificação em Fluxo de Dados

- Classificação efetiva requer::
  - Atualização do modelo de classificação à medida que o fluxo evolui.
    - Considerar limitação de recursos: memória, tempo and dados rotulados.



Questão de Pesquisa

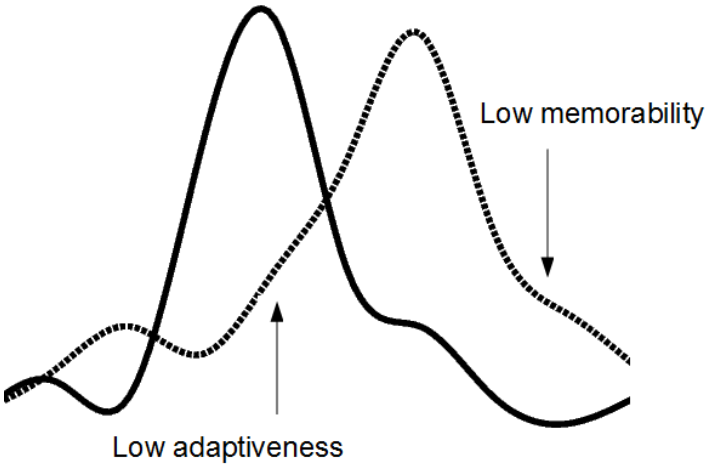
Como lidar com mudanças de conceito?

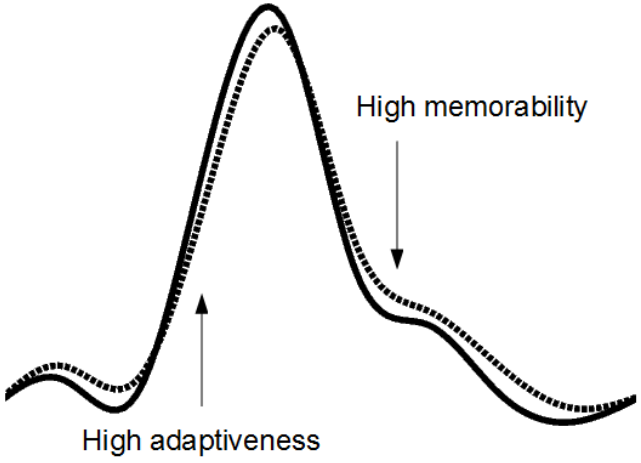
Modelo de Classificação

- Modelos de classificação compostos por regras de associação.
  - $\{x \rightarrow y\}$, onde $x \in X$ e $y \in Y$
- Atualização eficiente à medida que o conjunto de treino evolui.
- Modelos são construidos sob demanda:
  - Para um dado $[x_i, *]$, regras $\{x \rightarrow y\}$ tal que $x \subseteq x_i$ são produzidas.
  - Previsão é realizada pela combinação destas regras.
- A cada instante é produzido um modelo $\mathcal{R}(x_i)$.

Lidando com Mudanças de Conceito

- Duas propriedades são necessárias para produzir modelos de classificação robustos a mudanças nos dados:
  - Adaptação:
    - Abilidade de adaptar às mudanças.
  - Memorização:
    - Capacidade de recuperar após mudanças.



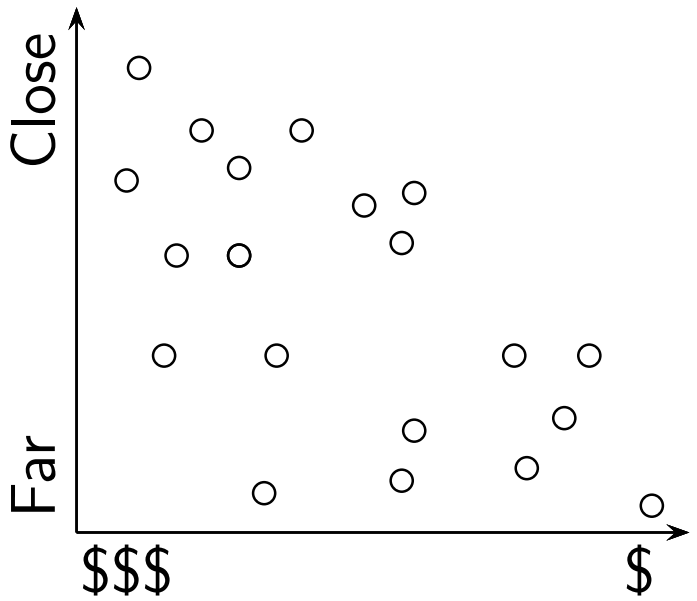


Dealing with Drifts

- Otimizando ambas propriedades leva a um problema de conflito de objetivo.
  - Otimizar adaptação pode prejudicar memorização, e vice-versa.

Eficiência Econômica

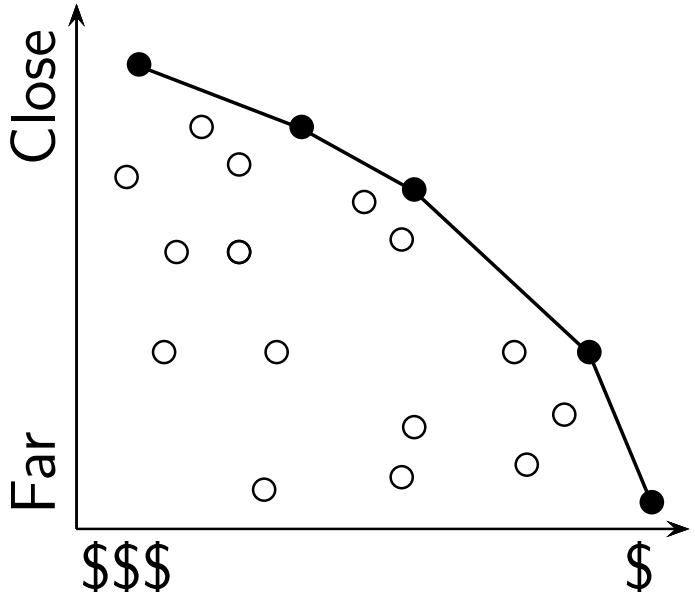
Exemplo: Hotéis em Petrópolis.



Eficência de Pareto

Fronteira de Pareto - Pontos Dominantes

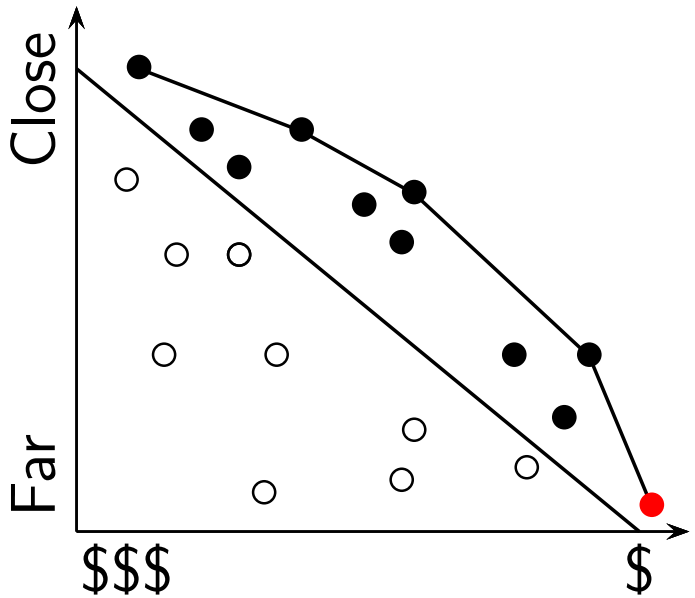
- $U_c(a) \geq U_c(b)$ e $U_d(a) \geq U_d(b)$
- $U_c(a) > U_c(b)$ ou $U_d(a) > U_d(b)$



Compensação — Princípio de Kaldor-Hicks

Região de Compensação:

- Utilidade total: $U(d_i) = U_m(d_i) + U_a(d_i)$
- Ponto base:
  $d^* = \{d_i \in \mathcal{P}_n | \forall d_j \in \mathcal{P}_n : U(d_i) \leq U(d_j)\}$



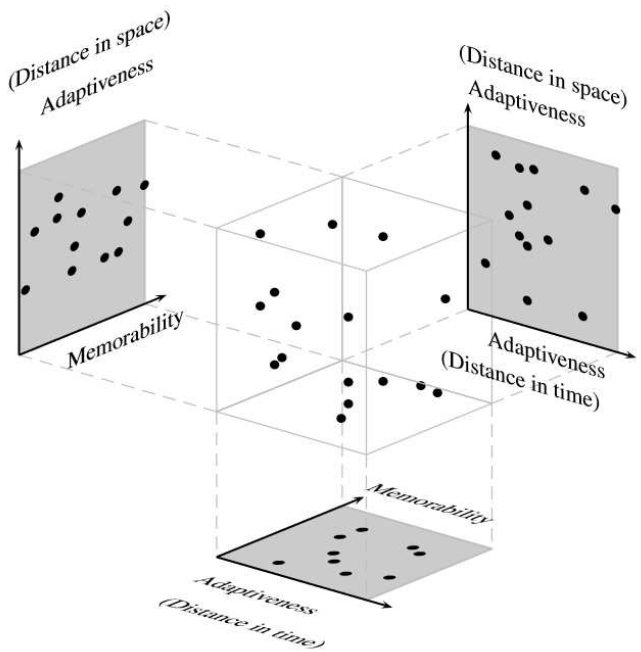
Medidas de Utilidade

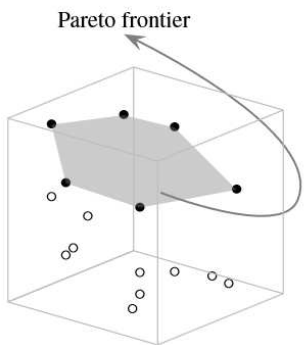
- Distância no espaço:
  - Similariade de cada instância de treino t_j em relação a nova instância t_n.
  - $U_s(t_j) = \frac{|\mathcal{R}(t_n) \cap \mathcal{R}(t_j)|}{|\mathcal{R}(t_n)|}$
- Distância no tempo:
  - Tempo de chegada de cada instância de treino t_j.
  - $U_t(t_j) = \frac{\gamma(t_j)}{\gamma(t_n)}$.
    - $\gamma(t_j)$ retorna o tempo em que a instância de treino t_j foi processada.
- Permutação aleatória das instâncias de treino:
  - $U_r(t_j) = \frac{\alpha(t_j)}{|\mathcal{D}_n|}$
    - $\alpha(t_j)$ retorna a posição de t_j na permutação.
    - $\mathcal{D}_n$ é o conjunto de treino a cada momento n.

Espaço de Utilidade

1. Coloque as instâncias de treino no espaço de utilidade.
2. Selecione as instâncias na Região de Eficiência:

    - Pareto-Efficience Selective Sampling (PESS)
    - Kaldor-Hicks Selective Sampling (KHSS).

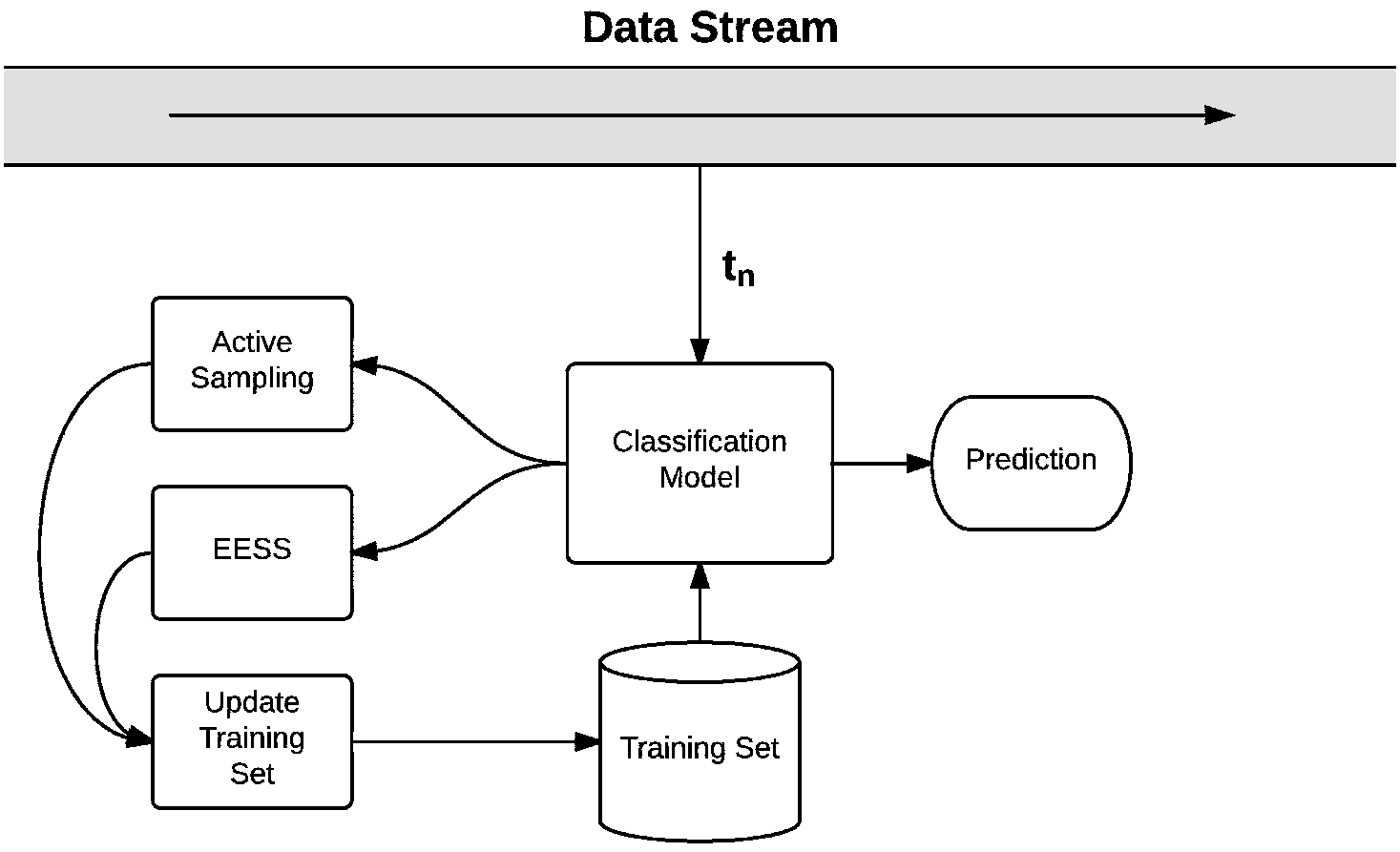




Redução do Esforço de Rotuação

- Amostragem Ativa Aleatória
  - Estratégia ingênua.
  - Simples para integrar.
  - Controle de Esforço de Rotulação: β.

Economically-Efficient Selective Sampling



Avaliação
Setup

- Interleaved Test-Then-Train.
- 1% do conjunto de dados provido como conjunto de treino.
- Ambiente de avaliação: Massive Online Analysis (MOA) framework.
- Baselines:

Algoritmo	Adaptação	Memorização
AC (KDD 2011)	Aprendizado Ativo	Classificador base
HAT (JMLR 2011)	ADWIN	Conjunto de Árvores
ILAC (SIGIR 2011)	Projeção de dados	Conjunto de treino incremental

Avaliação

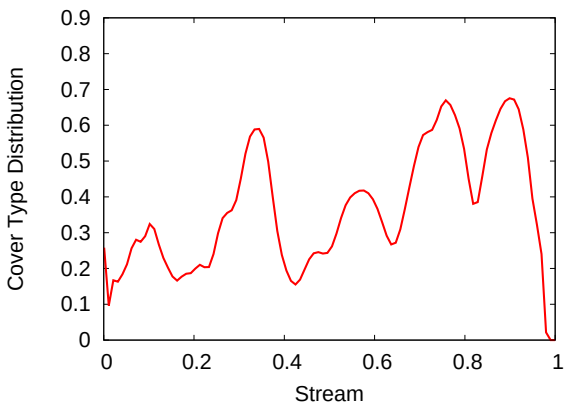
- Métricas:
  - Erro Quadrático Médio.
  - Esforço de Rotuação: 10%; **25%**; 50%; 75% and 100%;
    - AC e EESS.
  - Tamanho do conjunto de treino.
  - RAM-Hours.
- Conjuntos de Dados:

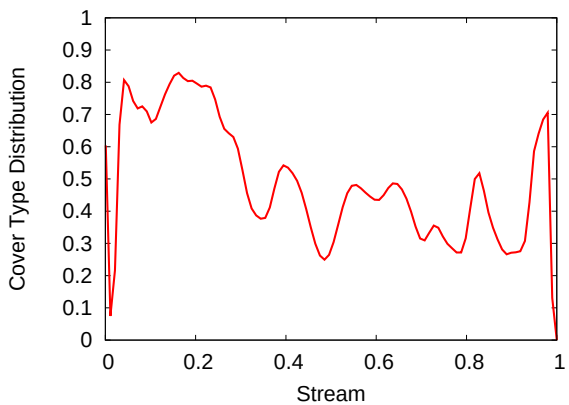
	Padrão de Mudança de Conceito			
	Repentino	Incremental	Gradual	Recorrente
Eleições Presidenciais 2010	-	X	X	-
Pessoa do Ano 2015	-	X	X	-
Copa do Mundo 2010 - Inglês	X	-	-	-
Copa do Mundo 2010 - PT	X	-	-	-
Tipo de Cobertura	X	-	X	X
Filtragem de Spam	X	-	X	X
Mão de Poker	-	-	X	X

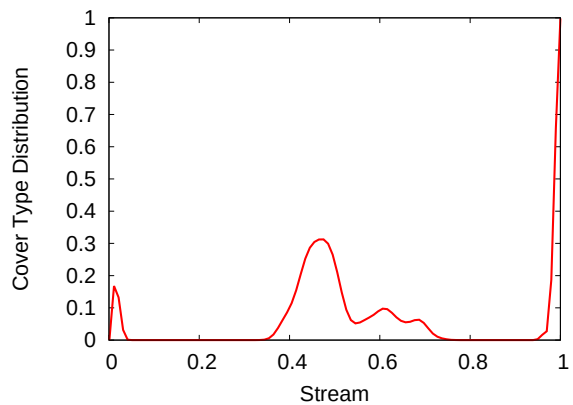
Evaluation

Tipo de cobertura de florestas

- Tipo de cobertura de florestas nos EUA.
- 581,102 instâncias com 54 variáveis e 7 classes;
- Mudanças de Conceito: Repentina, Gradual, Recorrent;



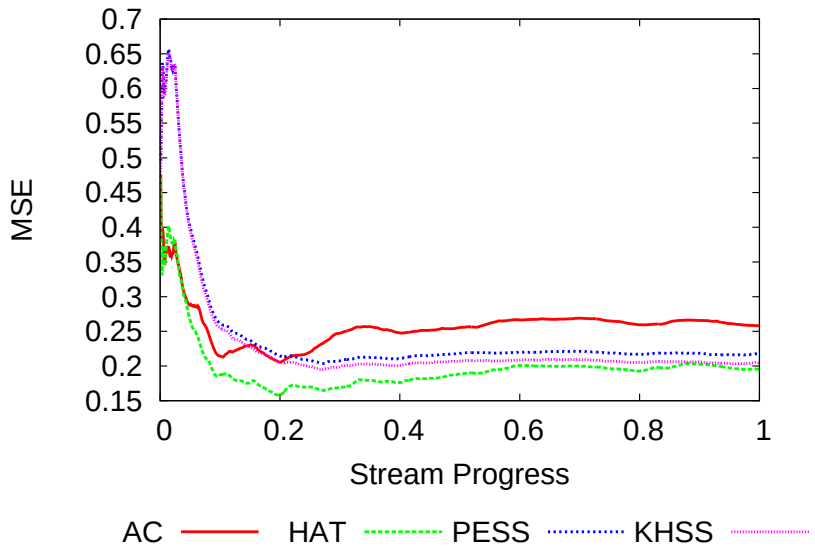


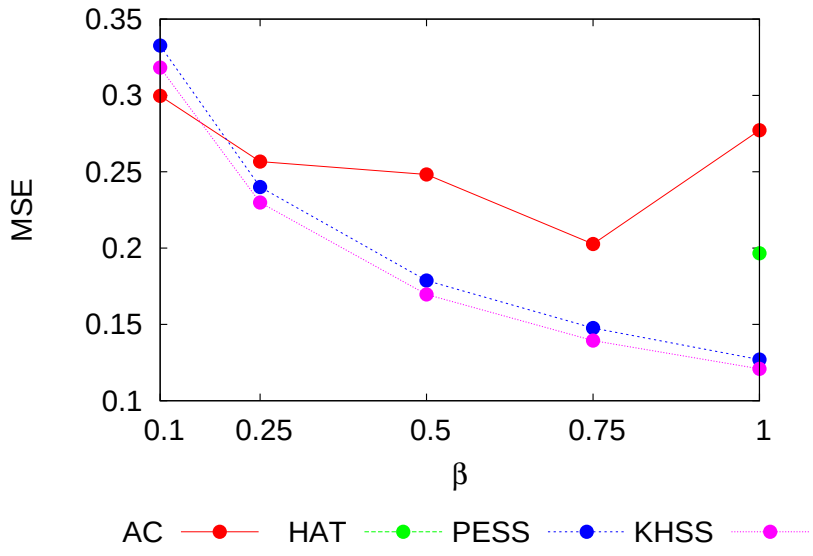


Evaluation

Tipo de cobertura de florestas

MSE and Esforço de Rotulação:

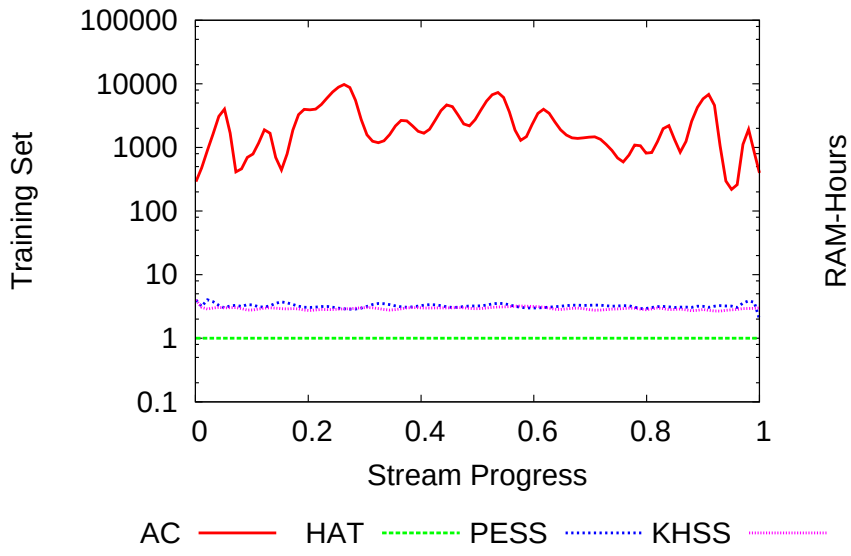


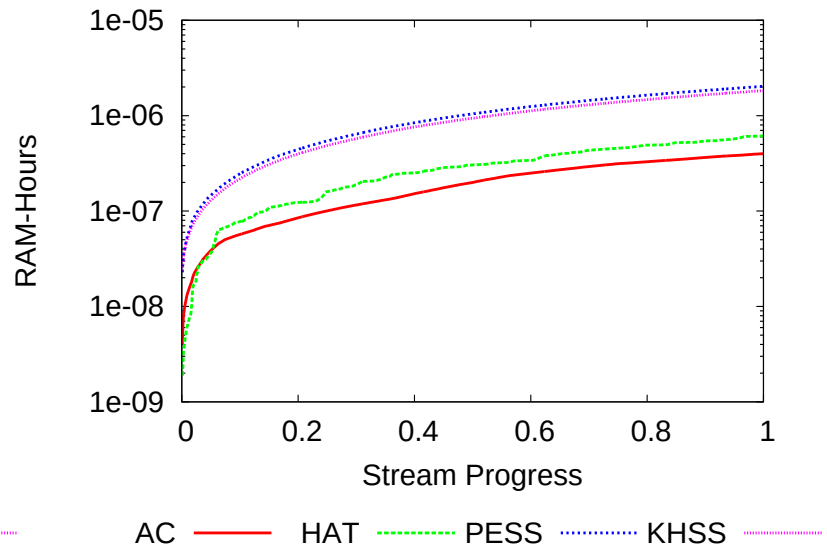


Evaluation

Tipo de cobertura de florestas

Tamanho do conjunto de treino e RAM-Hours





Conclusões

- Análise de Fluxos de Dados.
  - Limitação de recursos.
  - Mudanas de Conceito.

Conclusões

- Análise de Fluxos de Dados.
  - Limitação de recursos.
  - Mudanas de Conceito.
- Eficência e Precisão.
  - Modelo de classificação incremental.
  - Adaptação e Memorização.
  - Eficência de Pareto e princípio de compensação.
  - Medidas de utilidade simple de computar.
  - Nossos algoritmos se mostraram robustos em diferentes cenários.

Conclusões

- Análise de Fluxos de Dados.
  - Limitação de recursos.
  - Mudanas de Conceito.
- Eficência e Precisão.
  - Modelo de classificação incremental.
  - Adaptação e Memorização.
  - Eficência de Pareto e princípio de compensação.
  - Medidas de utilidade simple de computar.
  - Nossos algoritmos se mostraram robustos em diferentes cenários.
- Trabalho futuros:
  - Outras medidas de utilidade.
  - Aplidar o nosso método para reudção de esforço de rotulação.
  - Explorar outros modelos de classificação.

Thank you!

robertolojr@dcc.ufmg.br