

# Uso de Redes de Função de Base Radial e Cadeias de Markov para detecção online de mudanças de conceito em fluxos contínuos de dados

---

**Discente:** Ruivaldo Neto

**Orientador:** Ricardo Rios

Universidade Federal da Bahia

Departamento de Ciência da Computação

Programa de Pós-Graduação em Ciência da Computação

Contato: [rneto@rneto.dev](mailto:rneto@rneto.dev)

16 de Dezembro de 2019

1. Introdução
2. Fundamentação Teórica
3. RBFChain
4. Experimentos
5. Conclusões e Trabalhos Futuros

# Introdução

---

- Avanços tecnológicos recentes contribuíram para o aumento do volume de dados produzidos por sistemas computacionais [8].
- Parte significativa desse volume é produzida na forma de **Fluxos Contínuos de Dados (FCDs)**: sequências **ininterruptas** e **potencialmente infinitas** de eventos [1].
- FCDs estão presentes em variados domínios de aplicação:
  - Análise do Mercado Financeiro;
  - Gestão de redes de telecomunicação;
  - Detecção de intrusos;
  - Monitoramento de tráfego; etc.

- Técnicas de **Aprendizado de Máquina (AM)** são utilizadas para extrair informações úteis desses grandes conjuntos de dados.
- Cenários com FCDs limitam a aplicação dessas técnicas, pois impõem restrições de tempo de resposta, de uso dos recursos computacionais e apresentam comportamento **não estacionário**.
- Em cenários **não estacionários**, o contexto do processo gerador e/ou a distribuição dos dados podem sofrer alterações ao longo do tempo.
- Essas mudanças, denominadas **mudanças de conceito** (*concept drifts*), podem ter impacto negativo nas técnicas de AM aplicadas.

- Inicialmente, realizava-se a atualização periódica dos modelos para mitigar os efeitos das mudanças de conceito.
- Pesquisadores propuseram métodos mais precisos e eficazes baseados em monitoramento.
- Os métodos propostos apresentam limitações ao serem aplicados em cenários com FCDs [1]:
  - Necessidade de rotulação;
  - Não atendem às restrições (tempo de resposta e uso de recursos).

- Visando superar essas limitações, este trabalho propõe um novo método de detecção de mudanças de conceito baseado em **Redes de Função de Base Radial (redes RBF)** e **Cadeias de Markov**, denominado **RBFCChain**.
- O método proposto se diferencia por detectar mudanças em **tempo de execução**, de forma computacionalmente **eficiente** e **independente de rótulos**.

# Fundamentação Teórica

---



- Fluxos Contínuos de Dados (FCDs) são sequências ininterruptas e potencialmente infinitas de eventos [1], que não podem ser armazenados e devem ser analisados em tempo de execução.
- Algoritmos de AM foram adaptados para atender a essas restrições [2], mas ainda são suscetíveis a mudanças de conceito.

# Mudança de Conceito

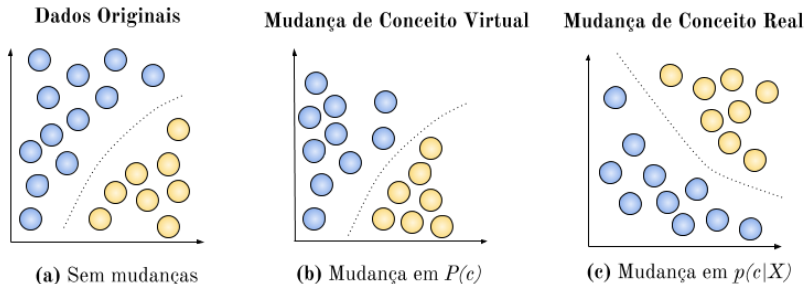
- **Mudanças de conceito** são alterações no contexto do processo e/ou na distribuição dos dados que podem impactar negativamente técnicas de AM.
- Podem ser definidas formalmente através da Teoria Bayesiana de Decisão [4]: sendo  $p_{t_0}$  e  $p_{t_1}$  as distribuições de probabilidades conjuntas nos instantes  $t_0$  e  $t_1$ , há mudança de conceito entre  $t_0$  e  $t_1$  se:

$$\exists X : p_{t_0}(X, c) \neq p_{t_1}(X, c) \quad (1)$$

# Mudança de Conceito

- São categorizadas como **Virtuais** ou **Reais** [5]:
  - **Mudanças Virtuais** são alterações na probabilidade a priori das classes,  $P(c)$ , e **não modificam** os resultados esperados.
  - **Mudanças Reais** são alterações na probabilidade a posteriori,  $p(c|X)$ , e **modificam** os resultados esperados.

# Mudança de Conceito



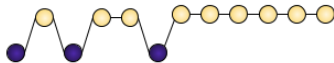
**Figura 1:** Mudança de Conceito Virtual vs. Mudança de Conceito Real.

# Mudança de Conceito

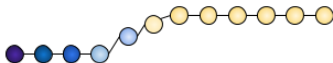
- Ocorrem de forma **abrupta**, **gradual**, **incremental** ou **recorrente** [7].



(a) Abrupta



(b) Gradual



(c) Incremental



(d) Recorrente

**Figura 2:** Padrões de ocorrência de Mudanças de Conceito.

# Algoritmos para Detecção de Mudança de Conceito

- Algoritmos de detecção se dividem em dois grupos, conforme a necessidade de rotulação dos dados [7]:
  - **Explícitos/Supervisionados**: **Dependem** da rotulação dos dados. Realizam a detecção a partir do monitoramento de medidas de performance como taxa de erro e acurácia.
  - **Implícitos/Não Supervisionados**: **Independem** da rotulação dos dados. Realizam a detecção através do monitoramento de características dos próprios dados ou de indicadores produzidos pelas técnicas de aprendizado aplicadas.

# Ferramenta: MOA

- O MOA é o principal framework para pesquisa com fluxos contínuos.
- Possui rotinas para produzir dados sintéticos e para avaliar métodos de detecção de mudança de conceito.

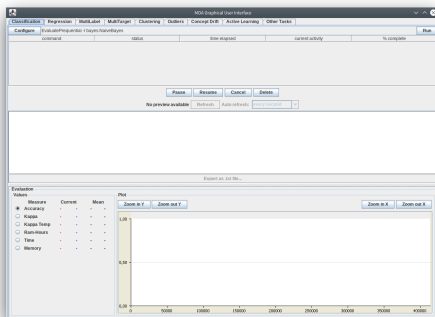


Figura 3: MOA - Tela Inicial.

# Mudança de Conceito - RBFChain

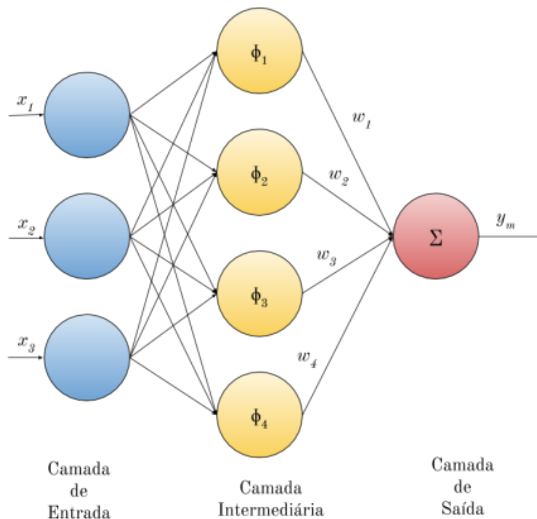
- O método proposto identifica mudanças **sob qualquer padrão de ocorrência**.
- É independente de rótulos, logo considera todas mudanças identificadas como **mudanças reais**.
- Foi implementado e validado através do MOA.



# Redes de Função de Base Radial

- **Redes de Função de Base Radial** são redes neurais cuja ativação é realizada através do cálculo da distância entre o evento e um centro definido [3].
- A arquitetura mais básica para redes RBF envolve três camadas:
  - **Entrada**: Recepciona os dados e encaminha para camada intermediária.
  - **Intermediária**: Composta por funções de ativação de base radial que atuam como neurônios.
  - **Saída**: Pondera os resultados da camada intermediária, agregando-os linearmente para compor a resposta final da rede.

# Redes de Função de Base Radial



**Figura 4:** Arquitetura RBF.

- O RBFCChain utiliza uma rede RBF adaptada, composta apenas pelas camadas inicial e intermediária.
- O processo de ativação da camada intermediária produz, implicitamente, grupos dos eventos recebidos.
- Mudanças de conceito são identificadas quando o grupo ativo desse agrupamento é alterado.

- **Cadeias de Markov** são processos estocásticos no qual a probabilidade do estado em um determinado período de tempo depende apenas do estado no período imediatamente anterior (Equação 2).

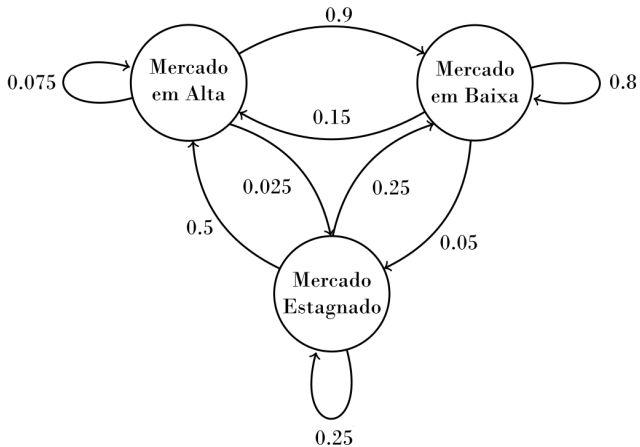
$$\mathbb{P}(X_t = s_j | X_{t-1} = s_i, \dots, X_0 = s_0) = \mathbb{P}(X_t = s_j | X_{t-1} = s_i) = p_{ij} \quad (2)$$

# Cadeias de Markov

- A Cadeia de Markov pode assumir os estados  $a_1, a_2, \dots, a_r$ , de tal modo que a probabilidade de transição de um estado  $a_i$  para um estado  $a_j$  seja  $P_{ij}$  (um valor dependente apenas de  $i$  e  $j$ );
- As probabilidades entre estados podem ser representadas por uma matriz estocástica (Equação 3):

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1r} \\ P_{21} & P_{22} & \dots & P_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ P_{r1} & P_{r2} & \dots & P_{rr} \end{bmatrix} \quad (3)$$

# Cadeias de Markov



**Figura 5:** Representação Gráfica: Cadeia de Markov com três estados.

- O RBFChain utiliza uma Cadeia de Markov para modelar o agrupamento criado na rede RBF adaptada.
- Os grupos formados representam os estados e as ativações de novos grupos, as transições.
- As transições são refletidas no modelo markoviano através do aumento da probabilidade correspondente e a diminuição proporcional das outras transições, respeitando a condição:  
 $0 \leq P_{ij} \leq 1$ .

- Pesquisa na literatura em busca de trabalhos que propõem métodos para identificação de mudanças de conceito em fluxos contínuos de dados, de forma online e independente de rótulos.
- Também foram estudadas técnicas que pudessem subsidiar o desenvolvimento de novos algoritmos que atendam a esses requisitos.

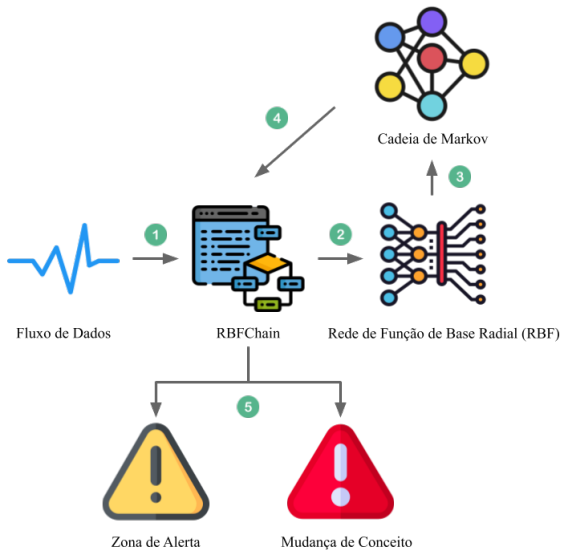


- Análise dos algoritmos Implícitos/Não Supervisionados da subcategoria Detecção de Novidade / Métodos de Agrupamento.
- Análise dos métodos para detecção de *Change Points* em séries temporais que atuam de forma online:
  - Modelos autoregressivos;
  - Séries com autosimilaridade e periodicidade.
- Análise da aplicação de algoritmos de agrupamento estáveis.
- Identificação de lacuna de pesquisa.

## RBFCChain

---

- O RBFCChain atua diretamente sobre o fluxo de dados e é composto por dois componentes principais: uma Rede de Função de Base Radial (RBF) adaptada e uma Cadeia de Markov.



**Figura 6:** Arquitetura RBFChain.

# Execução de exemplo

- $S = 0.11, 0.12, 0.13, 0.33, 0.34, 0.45, 0.6, 0.33, 0.25, 0.14, 0.11, 0.15$
- $\sigma = 3, \lambda = 0.8, \alpha = 0.25, \delta = 0.5$

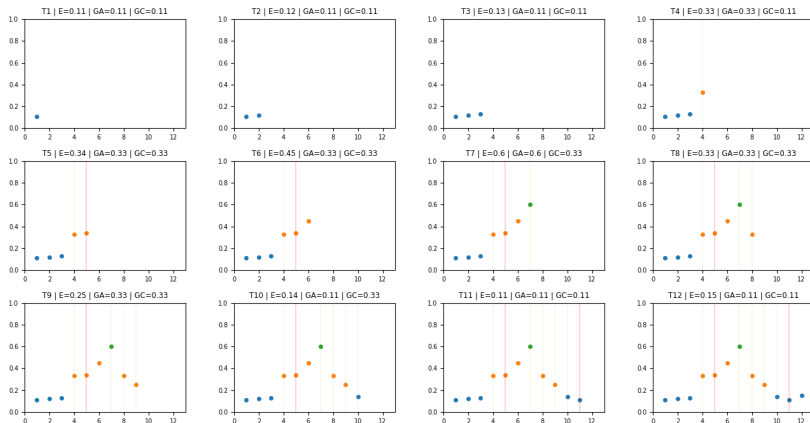


Figura 7: Execução de exemplo do RBFChain.

# Execução de exemplo

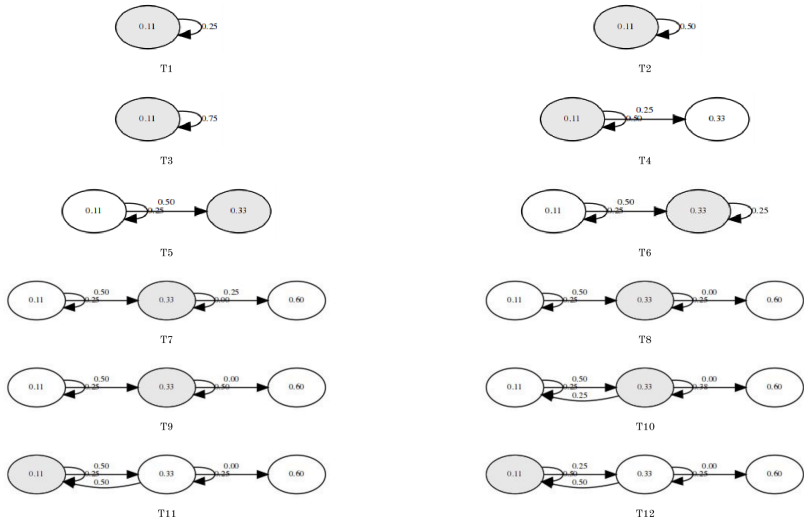


Figura 8: Evolução do modelo markoviano.

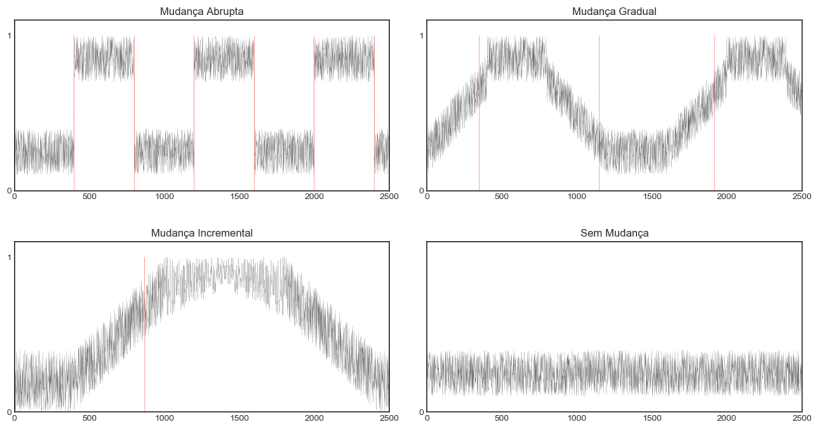
# Experimentos

---

- O primeiro experimento utilizou dados sintéticos produzidos por versões adaptadas das classes geradoras do MOA.
- Foram utilizados 4 conjuntos de dados, com 2.500 eventos cada.
- Os eventos possuem valores entre 0 e 1, com adição de um ruído randômico entre  $[-0.1, 0.2]$ .
- Cada conjunto de dados pode representar até 2 conceitos distintos.
- Cada conceito é composto por 400 eventos.



# Dados Sintéticos

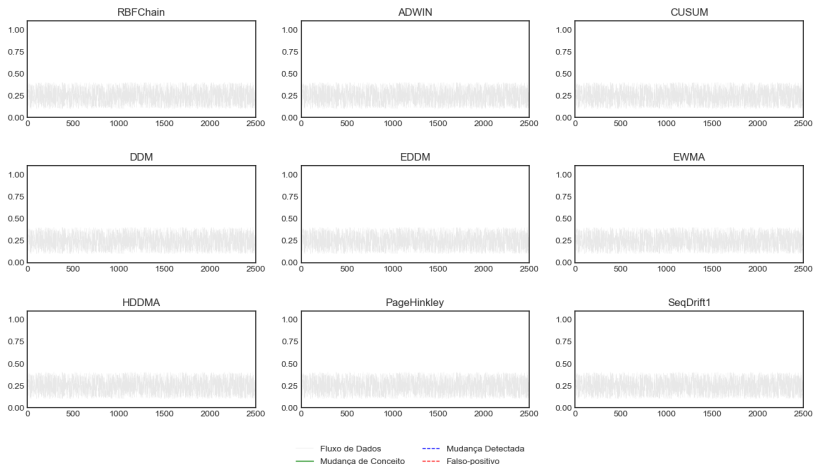


**Figura 9:** Representação gráfica dos conjuntos de dados sintéticos.

**Tabela 1:** Métricas utilizadas na avaliação com dados sintéticos.

Métrica	Observação
TP	<b>Tempo de Processamento</b> por instância (média em seg.).
MR	<b>Mudanças Reais</b> existentes no conjunto de dados.
VP	<b>Verdadeiro Positivo.</b> Quantidade de detecções corretas.
FP	<b>Falso Positivo.</b> Quantidade de detecções errôneas.
ATR	<b>Atraso de detecção.</b> Quantidade média de instâncias até a detecção.

# Dados Sintéticos - Sem mudanças de conceito



**Figura 10:** Comportamento dos algoritmos para o conjunto de dados sem mudanças de conceito.

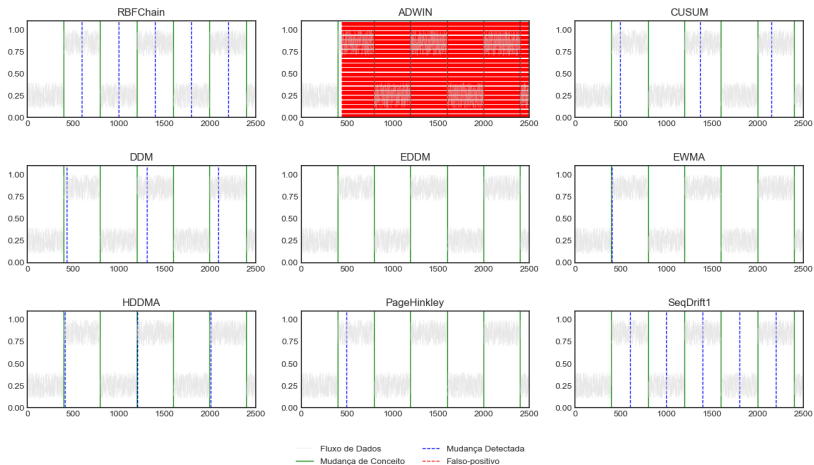
## Dados Sintéticos - Sem mudanças de conceito

**Tabela 2:** Resultados dos algoritmos para o conjunto de dados sem mudanças de conceito.

Algoritmo	TP	MR	VP	FP	ATR
RBFChain	0.013	0	0	0	—
ADWIN	0.025	0	0	0	—
CUSUM	0.016	0	0	0	—
DDM	0.014	0	0	0	—
EDDM	0.013	0	0	0	—
EWMA	0.014	0	0	0	—
HDDMA	0.017	0	0	0	—
PageHinkley	0.015	0	0	0	—
SeqDrift1	0.017	0	0	0	—

- Todos algoritmos testados demonstraram tolerância a ruídos e não indicaram nenhum falso positivo.
- RBFChain obteve a melhor média em tempo de processamento, juntamente com o EDDM.

# Dados Sintéticos - Mudanças Abruptas



**Figura 11:** Comportamento dos algoritmos para o conjunto de dados com mudanças de conceito abruptas.

## Dados Sintéticos - Mudanças Abruptas

**Tabela 3:** Resultados dos algoritmos para o conjunto de dados com mudanças de conceito abruptas.

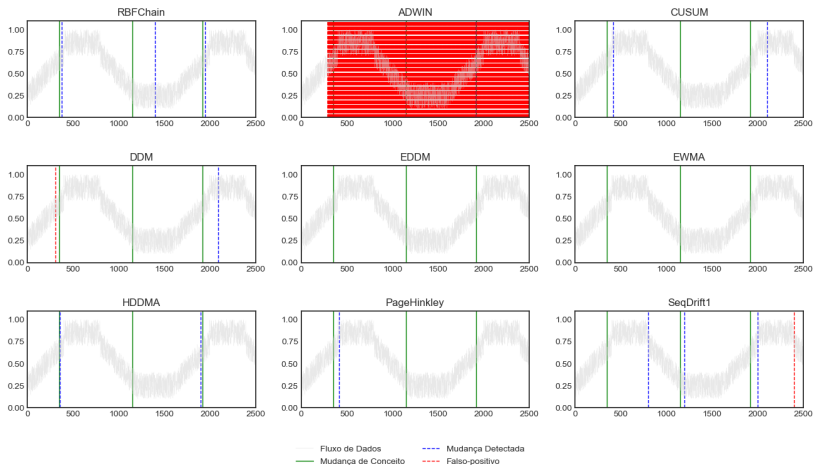
Algoritmo	TP	MR	VP	FP	ATR
RBFChain	0.015	6	5	0	166.67
ADWIN	0.016	6	6	2046	9.00
CUSUM	0.021	6	3	0	68.83
DDM	0.015	6	3	0	38.83
EDDM	0.013	6	0	0	—
EWMA	0.014	6	1	0	1.00
HDDMA	0.014	6	3	0	10.00
PageHinkley	0.015	6	1	0	16.17
SeqDrift1	0.015	6	5	0	167.50

## Dados Sintéticos - Mudanças Abruptas

- RBFChain e SeqDrift1 apresentaram as melhores acurácias, identificando 5 das 6 mudanças existentes, sem produzir nenhum falso positivo.
- RBFChain apresentou o terceiro melhor tempo de processamento (TP), com uma pequena diferença para o primeiro.
- ADWIN se mostrou hipersensível.



# Dados Sintéticos - Mudanças Graduais



**Figura 12:** Comportamento dos algoritmos para o conjunto de dados com mudanças de conceito graduais.

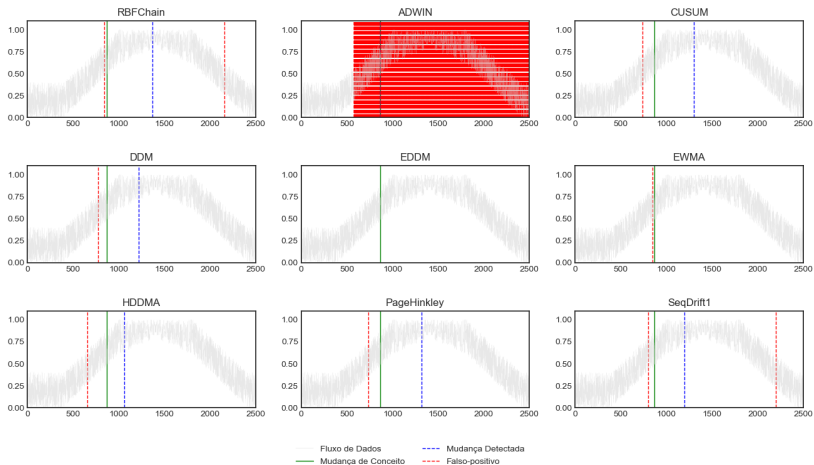
## Dados Sintéticos - Mudanças Graduais

**Tabela 4:** Resultados dos algoritmos para o conjunto de dados com mudanças de conceito graduais.

Algoritmo	TP	MR	VP	FP	ATR
RBFChain	0.011	3	3	0	101.00
ADWIN	0.020	3	3	2209	1.00
CUSUM	0.014	3	2	0	83.33
DDM	0.014	3	1	1	58.33
EDDM	0.013	3	0	0	—
EWMA	0.015	3	0	0	—
HDDMA	0.014	3	2	0	100.67
PageHinkley	0.014	3	1	0	22.33
SeqDrift1	0.015	3	3	1	194.33

- RBFChain obteve a melhor acurácia, identificando todas as três mudanças de conceito, sem produzir falso positivos.
- Algoritmo SeqDrift1 apresentou a segunda melhor acurácia, pois também detectou as três mudanças corretamente, entretanto, apresentou um falso positivo e uma taxa de atraso significativamente maior.
- RBFChain também apresentou o melhor tempo de processamento.

# Dados Sintéticos - Mudanças Incrementais



**Figura 13:** Comportamento dos algoritmos para o conjunto de dados com mudanças de conceito incrementais.

**Tabela 5:** Resultados dos algoritmos para o conjunto de dados com mudanças de conceito incrementais.

Algoritmo	TP	MR	VP	FP	ATR
RBFChain	0.020	1	1	2	501.00
ADWIN	0.017	1	1	1923	1.00
CUSUM	0.014	1	1	1	434.00
DDM	0.014	1	1	1	349.00
EDDM	0.013	1	0	0	—
EWMA	0.016	1	0	1	—
HDDMA	0.014	1	1	1	213.00
PageHinkley	0.014	1	1	1	449.00
SeqDrift1	0.016	1	1	2	331.00

## Dados Sintéticos - Mudanças Incrementais

- A mudança de conceito incremental representa uma dificuldade, pois todos algoritmos que detectaram a mudança existente também produziram falso positivos.
- RBFChain e SeqDrift1, que apresentaram os melhores resultados nos testes anteriores, tiveram a pior acurácia, pois emitiram dois falsos positivo cada.
- Teste considerado inconclusivo, ressaltando a necessidade de uma investigação mais detalhada sobre a detecção de mudanças de conceito incrementais.

## Dados Reais - Identificação de fixações e sacadas

- Dados de monitoramento ocular têm sido utilizados por uma quantidade significativa de pesquisas, em diferentes áreas do conhecimento.
- Os dados brutos coletados precisam ser identificados em eventos, como fixações e sacadas, para serem analisados.
- O processo de identificação é realizado por algoritmos. Entretanto, nenhum algoritmo existente na literatura permite identificar eventos em tempo de execução.
- Para superar esta limitação, empregamos o RBFChain na identificação de fixações e sacadas.

## Dados Reais - Identificação de fixações e sacadas



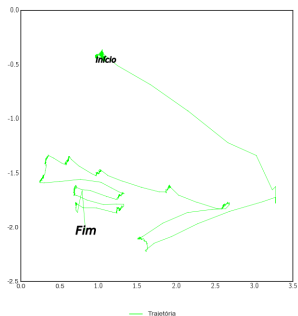
**Figura 14:** Exemplo de identificação de fixações e sacadas. Fonte: [6].



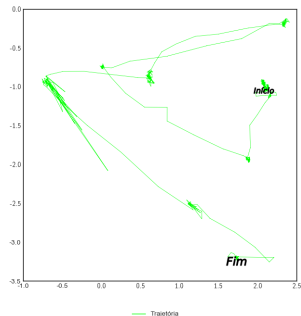
## Dados Reais - Identificação de fixações e sacadas

- O experimento utilizou dados de dois macacos-prego (Dede e Juju) produzidos e cedidos pelo Instituto do Cérebro (UFRN).
- Cada conjunto de dados possui 6.200 eventos, que indicam a localização do olhar ao longo do tempo  $(x, y)$ .
- O RBFChain foi ligeiramente adaptado para analisar a alternância (sacadas) e a continuidade (fixações) dos conceitos.
- Os resultados foram validados através de métricas de classificação, utilizando os resultados do algoritmo ClusterFix [6] como rótulos.

# Dados Reais - Identificação de fixações e sacadas - Trajetórias



**Figura 15:** Trajetória *Dede*.

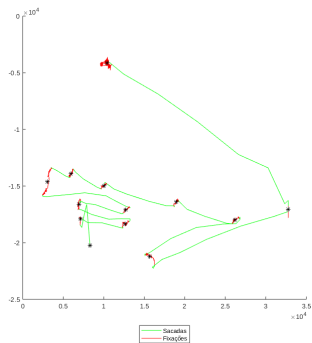


**Figura 16:** Trajetória *Juju*.

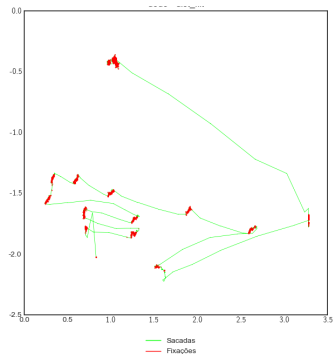
**Tabela 6:** Métricas utilizadas na avaliação com dados reais.

Métrica	Observação
QP	<b>Quantidade de Pontos</b> analisados.
AC	<b>Acurácia.</b> Fração de fixações e sacadas identificadas corretamente.
PR	<b>Precisão.</b> Fração das fixações identificadas pelo algoritmo corretamente.
RE	<b>Recall.</b> Fração das fixações existentes (rotuladas) que também foram identificadas pelo algoritmo.

# Dados Reais - Identificação de fixações e sacadas - Dede



**Figura 17:** ClusterFix - *Dede*.



**Figura 18:** RBFChain - *Dede*.

**Tabela 7:** Resultados para o conjunto de dados *Dede*.

QT	AC	PR	RE
6.200	0.87	0.98	0.88

- 87% das fixações e sacadas identificadas pelo RBFChain tiveram a mesma classificação pelo ClusterFix.
- 98% das fixações identificadas pelo RBFChain tiveram a mesma classificação pelo ClusterFix.
- 88% das fixações identificadas pelo ClusterFix também foram identificadas pelo RBFChain.

# Dados Reais - Identificação de fixações e sacadas - Juju

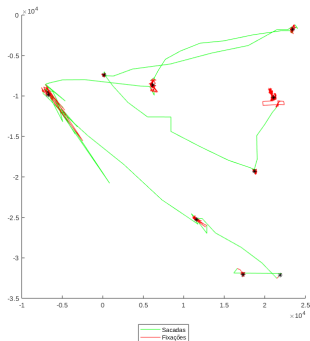


Figura 19: ClusterFix - *Juju*.

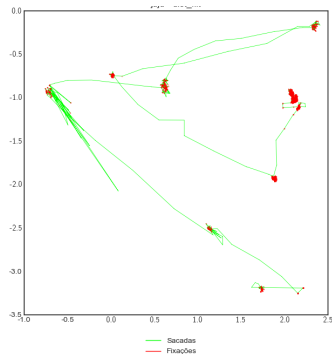


Figura 20: RBFChain - *Juju*.

**Tabela 8:** Resultados para o conjunto de dados *Juju*.

QT	AC	PR	RE
6.200	0.82	0.98	0.83

- 82% das fixações e sacadas identificadas pelo RBFChain tiveram a mesma classificação pelo ClusterFix.
- 98% das fixações identificadas pelo RBFChain tiveram a mesma classificação pelo ClusterFix.
- 83% das fixações identificadas pelo ClusterFix também foram identificadas pelo RBFChain.

## **Conclusões e Trabalhos Futuros**

---



- Novo método de detecção de mudanças de conceito, capaz de detectar mudanças em tempo de execução, de forma computacionalmente eficiente e independente de rótulos;
- Novo método para identificação de fixações e sacadas em tempo de execução, com acurácia equivalente ao estado da arte.

- Desenvolvimento de novas estratégias para o cálculo dos parâmetros.
- Criação de novas bases de dados experimentais.
- Investigação aprofundada sobre a detecção de mudanças incrementais.



C. C. Aggarwal.

**Data Streams: Models and Algorithms (Advances in Database Systems).**

Springer-Verlag, Berlin, Heidelberg, 2006.



C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu.

**A framework for clustering evolving data streams.**

In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, VLDB '03, pages 81–92. VLDB Endowment, 2003.



A. Braga, A. C. Carvalho, and T. B. Ludermir.

**Redes Neurais Artificiais: Teoria e aplicações, volume 2.**

LTC Editora, 2007.



R. O. Duda, P. E. Hart, and D. G. Stork.

**Pattern Classification (2Nd Edition).**

Wiley-Interscience, New York, NY, USA, 2000.



J. a. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia.

**A survey on concept drift adaptation.**

*ACM Comput. Surv.*, 46(4):44:1–44:37, Mar. 2014.



S. D. König and E. A. Buffalo.

**A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds.**

*Journal of Neuroscience Methods*, 227:121 – 131, 2014.



I. Zliobaite.

**Learning under concept drift: an overview.**

*CoRR*, abs/1010.4784, 2010.



M. Zwolenski and L. Weatherill.

**The digital universe rich data and the increasing value of the internet of things.**

*Australian Journal of Telecommunications and the Digital Economy*,  
2, 10 2014.