Proposta de uma abordagem para a 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

14 de Junho de 2019

Roteiro

1. Introdução
2. Revisão Bibliográfica
3. Plano de Pesquisa
4. Experimentos Iniciais
5. Conclusão

Introdução

Contexto e Motivação

- Avanços tecnológicos recentes favoreceram o aumento no volume de dados produzidos por sistemas computacionais.

Contexto e Motivação

- Avanços tecnológicos recentes favoreceram o aumento no volume de dados produzidos por sistemas computacionais.
- Muitos desses dados são produzidos na forma de sequências ininterruptas e potencialmente infinitas, denominadas Fluxos Contínuos de Dados (FCDs) [2].

Contexto e Motivação

- Para extrair informações úteis desses grandes conjuntos, técnicas da área de Aprendizado de Máquina (AM) têm sido aplicadas.

Contexto e Motivação

- Para extrair informações úteis desses grandes conjuntos, técnicas da área de Aprendizado de Máquina (AM) têm sido aplicadas.
- Contudo, além das restrições impostas pelos fluxos contínuos, estas técnicas também devem lidar com alterações no contexto do processo gerador do fluxo e/ou na distribuição dos dados.

Contexto e Motivação

- Para extrair informações úteis desses grandes conjuntos, técnicas da área de Aprendizado de Máquina (AM) têm sido aplicadas.
- Contudo, além das restrições impostas pelos fluxos contínuos, estas técnicas também devem lidar com alterações no contexto do processo gerador do fluxo e/ou na distribuição dos dados.
- Estas alterações são denominadas Mudanças de Conceito e podem impactar a acurácia do modelo.

Contexto e Motivação

- Inicialmente, a atualização periódica do modelo foi utilizada como técnica para mitigar a perda causada por essas mudanças.

Contexto e Motivação

- Inicialmente, a atualização periódica do modelo foi utilizada como técnica para mitigar a perda causada por essas mudanças.
- Com o avanço da pesquisa, métodos de detecção de mudança baseados em monitoramento foram propostos.

Contexto e Motivação

- Os métodos existentes na literatura apresentam limitações ao serem aplicados em cenários com fluxos contínuos de dados.

Contexto e Motivação

- Os métodos existentes na literatura apresentam limitações ao serem aplicados em cenários com fluxos contínuos de dados.
- As principais limitações são: a necessidade de rótulação e/ou tempo de resposta elevado.

Contexto e Motivação

- Os métodos existentes na literatura apresentam limitações ao serem aplicados em cenários com fluxos contínuos de dados.
- As principais limitações são: a necessidade de rótulação e/ou tempo de resposta elevado.
- Visando mitigá-las, este trabalho discute uma abordagem baseada em Redes de Função de Base Radial (RBF) para detecção de mudanças de conceito em FCDs.

Hipótese

"A aplicação de Redes de Função de Base Radial em fluxos contínuos de dados permite a detecção de mudanças de conceito em tempo de execução, de forma computacionalmente eficiente e independente de rótulos."

Hipótese

- O objetivo deste trabalho é a validação da hipótese. Para tanto, um novo método de detecção, baseado em redes RBF, será desenvolvido.

Hipótese

- O objetivo deste trabalho é a validação da hipótese. Para tanto, um novo método de detecção, baseado em redes RBF, será desenvolvido.
- O método será validado através de comparações com o estado da arte.

Hipótese

- O objetivo deste trabalho é a validação da hipótese. Para tanto, um novo método de detecção, baseado em redes RBF, será desenvolvido.
- O método será validado através de comparações com o estado da arte.
- Dois conjuntos de dados serão utilizados nos experimentos, um sintético e outro oriundo de uma aplicação do mundo real.

Revisão Bibliográfica

Fluxos Contínuos de Dados e Aprendizado de Máquina

- Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências ininterruptas e potencialmente infinitas de eventos [2]

Fluxos Contínuos de Dados e Aprendizado de Máquina

- Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências ininterruptas e potencialmente infinitas de eventos [2]
- Estes fluxos não podem ser armazenados em sua totalidade e, por serem de alta frequência, devem ser analisados em tempo real.

Fluxos Contínuos de Dados e Aprendizado de Máquina

- Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências ininterruptas e potencialmente infinitas de eventos [2]
- Estes fluxos não podem ser armazenados em sua totalidade e, por serem de alta frequência, devem ser analisados em tempo real.
- Algoritmos supervisionados [6, 5, 12, 4, 8] e não-supervisionados [3, 1, 11] foram adaptados para atenderem a essas restrições.

Fluxos Contínuos de Dados e Aprendizado de Máquina

- Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências ininterruptas e potencialmente infinitas de eventos [2]
- Estes fluxos não podem ser armazenados em sua totalidade e, por serem de alta frequência, devem ser analisados em tempo real.
- Algoritmos supervisionados [6, 5, 12, 4, 8] e não-supervisionados [3, 1, 11] foram adaptados para atenderem a essas restrições.
- Entretanto, o contexto do processo gerador ou a distribuição dos dados gerados podem sofrer alterações.

Fluxos Contínuos de Dados e Aprendizado de Máquina

- Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências ininterruptas e potencialmente infinitas de eventos [2]
- Estes fluxos não podem ser armazenados em sua totalidade e, por serem de alta frequência, devem ser analisados em tempo real.
- Algoritmos supervisionados [6, 5, 12, 4, 8] e não-supervisionados [3, 1, 11] foram adaptados para atenderem a essas restrições.
- Entretanto, o contexto do processo gerador ou a distribuição dos dados gerados podem sofrer alterações.
- Estas alterações são denominadas mudanças de conceito e podem impactar as técnicas de aprendizado aplicadas.

Mudança de Conceito

- A Teoria Bayesiana de Decisão [7] é comumente utilizada para descrever a tarefa de classificação.

Mudança de Conceito

- A Teoria Bayesiana de Decisão [7] é comumente utilizada para descrever a tarefa de classificação.
- Baseando-se nesta teoria e considerando que p_{t_0} e p_{t_1} denotam as distribuições de probabilidades conjuntas nos instantes t_0 e t_1, é possível afirmar que há mudança de conceito entre os instantes t_0 e t_1 se:

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

Mudança de Conceito

- A Teoria Bayesiana de Decisão [7] é comumente utilizada para descrever a tarefa de classificação.
- Baseando-se nesta teoria e considerando que p_{t_0} e p_{t_1} denotam as distribuições de probabilidades conjuntas nos instantes t_0 e t_1, é possível afirmar que há mudança de conceito entre os instantes t_0 e t_1 se:

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

- Um conjunto de dados possui resultados esperados legítimos em t_0, mas este mesmo conjunto passa a ter resultados esperados diferentes, também legítimos, em t_1 [10].

Mudança de Conceito

- As mudanças de conceito podem ser categorizadas como virtuais ou reais [9].

Mudança de Conceito

- As mudanças de conceito podem ser categorizadas como virtuais ou reais [9].
- As mudanças virtuais são causadas por alterações na probabilidade a priori das classes, $P(c)$, e não alteram os conceitos-alvo.

Mudança de Conceito

- As mudanças de conceito podem ser categorizadas como virtuais ou reais [9].
- As mudanças virtuais são causadas por alterações na probabilidade a priori das classes, $P(c)$, e não alteram os conceitos-alvo.
- Enquanto que as mudanças de conceito reais surgem a partir de alterações na probabilidade a posteriori, $p(c|X)$, e modificam os resultados esperados.

Mudança de Conceito

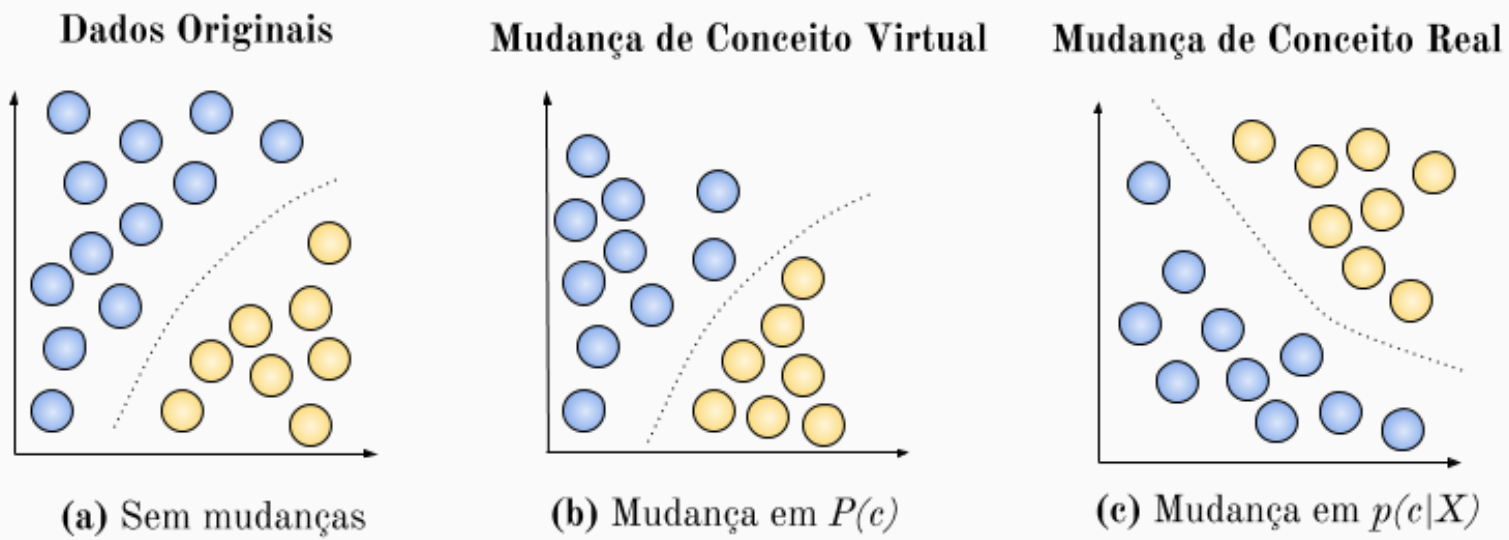

Figura 1: Mudança de Conceito Virtual vs. Mudança de Conceito Real

Mudança de Conceito

- As mudanças de conceito podem ocorrer de forma abrupta, gradual, incremental ou recorrente.

Mudança de Conceito

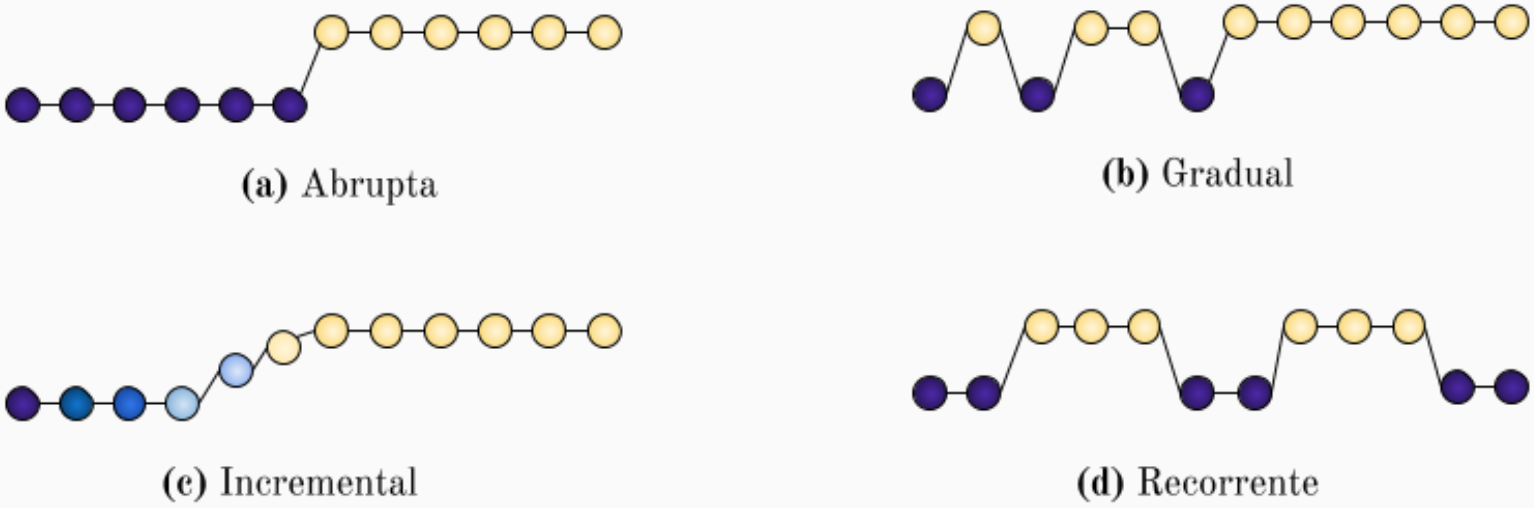

Figura 2: Padrões de ocorrência de Mudanças de Conceito

Mudança de Conceito

- O fenômeno mudança de conceito tem sido estudado em diferentes comunidades de pesquisa, incluindo Mineração de Dados, Aprendizado de Máquina, Estatística e Recuperação de Informação [13]. Contudo, o tema apresenta diferentes nomeclaturas em cada comunidade.

Mudança de Conceito

Tabela 1: Terminologia - Mudança de Conceito [13]

Área	Termos
Mineração de Dados	Mudança de Conceito
Aprendizado de Máquina	Mudança de Conceito, Mudança de Covariável
Computação Evolucionária	Ambiente Evolutivo, Ambiente em Mudança
IA e Robótica	Ambiente Dinâmico
Estatística, Séries Temporais	Não Estacionário
Recuperação de Informação	Evolução Temporal

Redes de Função de Base Radial

- A

Redes de Função de Base Radial

- A
- B

Redes de Função de Base Radial

- A
- B
- C

Redes de Função de Base Radial

- A
- B
- C
- D

Trabalhos Relacionados

- A

Trabalhos Relacionados

- A
- B

Trabalhos Relacionados

- A
- B
- C

Trabalhos Relacionados

- A
- B
- C
- D

Plano de Pesquisa

Descrição do Problema

- A

Descrição do Problema

- A
- B

Descrição do Problema

- A
- B
- C

Descrição do Problema

- A
- B
- C
- D

Atividades de Pesquisa

- A

Atividades de Pesquisa

- A
- B

Atividades de Pesquisa

- A
- B
- C

Atividades de Pesquisa

- A
- B
- C
- D

Experimentos Iniciais

Configuração dos Experimentos

- A

Configuração dos Experimentos

- A
- B

Configuração dos Experimentos

- A
- B
- C

Configuração dos Experimentos

- A
- B
- C
- D

Critérios de Avaliação

- A

Critérios de Avaliação

- A
- B

Critérios de Avaliação

- A
- B
- C

Critérios de Avaliação

- A
- B
- C
- D

Pettitt

- A

Pettitt

- A
- B

Pettitt

- A
- B
- C

Pettitt

- A
- B
- C
- D

RBFDriftDetector

- A

RBFDriftDetector

- A
- B

RBFDriftDetector

- A
- B
- C

RBFDriftDetector

- A
- B
- C
- D

Conclusão

Conclusão

- A

Conclusão

- A
- B

Conclusão

- A
- B
- C

Conclusão

- A
- B
- C
- D

Trabalhos Futuros

- A

Trabalhos Futuros

- A
- B

Trabalhos Futuros

- A
- B
- C

Trabalhos Futuros

- A
- B
- C
- D

Referências i


M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler.
Streamkm++: A clustering algorithm for data streams.
J. Exp. Algorithmics, 17:2.4:2.1–2.4:2.30, May 2012.


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.

Referências ii


C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu.
On demand classification of data streams.
In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 503–508, New York, NY, USA, 2004. ACM.


A. Bifet, B. Pfahringer, J. Read, and G. Holmes.
Efficient data stream classification via probabilistic adaptive windows.
In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 801–806, New York, NY, USA, 2013. ACM.

Referências iii


P. Domingos and G. Hulten.
Mining high-speed data streams.
In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 71–80, New York, NY, USA, 2000. ACM.


R. O. Duda, P. E. Hart, and D. G. Stork.
Pattern Classification (2Nd Edition).
Wiley-Interscience, New York, NY, USA, 2000.


J. a. Gama, R. Rocha, and P. Medas.
Accurate decision trees for mining high-speed data streams.
In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 523–528, New York, NY, USA, 2003. ACM.

Referências iv


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.


J. Z. Kolter and M. A. Maloof.
Dynamic weighted majority: An ensemble method for drifting concepts.
J. Mach. Learn. Res., 8:2755–2790, Dec. 2007.


P. Kranen, I. Assent, C. Baldauf, and T. Seidl.
The clustree: Indexing micro-clusters for anytime stream mining.
Knowl. Inf. Syst., 29(2):249–272, Nov. 2011.

Referências v

H. Wang, W. Fan, P. S. Yu, and J. Han.
Mining concept-drifting data streams using ensemble classifiers.
In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 226–235, New York, NY, USA, 2003. ACM.

I. Zliobaite.
Learning under concept drift: an overview.
CoRR, abs/1010.4784, 2010.