Universidade Federal da Bahia
Instituto de Matemática

Programa de Pós-Graduação em Ciência da Computação

DETECÇÃO DE MUDANÇAS DE CONCEITO EM FLUXOS DE DADOS NÃO ESTACIONÁRIOS

Ruivaldo Azevedo Lobão Neto

QUALIFICAÇÃO DE MESTRADO

Salvador
19 de Julho de 2018

RUIVALDO AZEVEDO LOBÃO NETO

DETECÇÃO DE MUDANÇAS DE CONCEITO EM FLUXOS DE DADOS NÃO ESTACIONÁRIOS

Esta Qualificação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: ***

Salvador
19 de Julho de 2018

RESUMO

O aprendizado a partir de fluxos de dados (aprendizagem incremental) tem crescido como foco de pesquisa, graças a existência de problemas práticos e desafios em aberto. Dentre estes, está a detecção de mudanças de conceito, fenômeno que ocorre quando a distribução dos dados é alterada, tornando o modelo vigente impreciso ou obsoleto. Neste trabalho, propomos uma nova técnica para detecção de mudanças de conceito.

Palavras-chave: Mudança de conceito, detecção de mudanças, aprendizagem adaptativa, fluxos de dados.

ABSTRACT

Learning from data streams (incremental learning) is increasing as a research focus, due to the existence of practical problems and open challenges. Among which, is the detection of concept drift, a phenomenon that happens when the data distribution is altered, making the model inaccurate or obsolete. In this work, we propose a novel technic to detect concept drifts.

Keywords: Concept drift, change detection, adaptive learning, data streams.

SUMÁRIO

Capítulo 1

REVISÃO BIBLIOGRÁFICA

1.0.1 Introdução

A extração de informações úteis a partir de grandes conjuntos de dados é uma tarefa desafiadora para os pesquisadores. Os algoritmos de aprendizagem de máquina baseados em fluxos de dados contínuos (FCDs) atuam em um contexto diferente dos algoritmos tradicionais, devido a natureza dinâmica das FCDs. Esses algoritmos devem se adaptar às constantes mudanças de distribuição dos dados, para não se tornarem imprecisos ou obsoletos.

Portanto, a atividade de Detecção de Novidades (DN) - *Concept Drift* - é essencial para o bom funcionamento dessas técnicas. A atividade de DN permite identificar o surgimento de novos conceitos e mudanças em conceitos existentes, possibilitando a atualização do modelo de decisão. Novas técnicas de aprendizado ativo têm sido exploradas com o objetivo de aprimorar o processo de classificação e identificação de mudanças de conceito.

1.0.2 Fluxos Contínuos de Dados

Fluxos Contínuos de Dados (FCDs) podem ser definidos como sequências contínuas de dados, de tamanho ilimitado, sem ordem definida e de alta frequência (BABCOCK et al., 2002). Novos algoritmos têm sido desenvolvidos para trabalhar com fluxos desse tipo, por exemplo: CLAM (AL-KHATEEB et al., 2012) e OLINDDA (SPINOSA; CARVALHO; GAMA, 2009). O desenvolvimento de algoritmos de aprendizado para esses cenários é uma tarefa custosa, pois devem lidar com sequências de dados geradas de forma contínua, em alta velocidade e cuja distribuição pode sofrer alterações ao longo do tempo (GAMA et al., 2014).

Os avanços recentes em hardware e software permitiram a aquisição de dados em maior escala, o que caracteriza ambientes dinâmicos, enquanto bases de dados tradicionais supõem cenários estacionários. O contínuo avanço das tecnologias, surgiram diversas aplicações do mundo real baseadas em fluxos contínuos de dados:

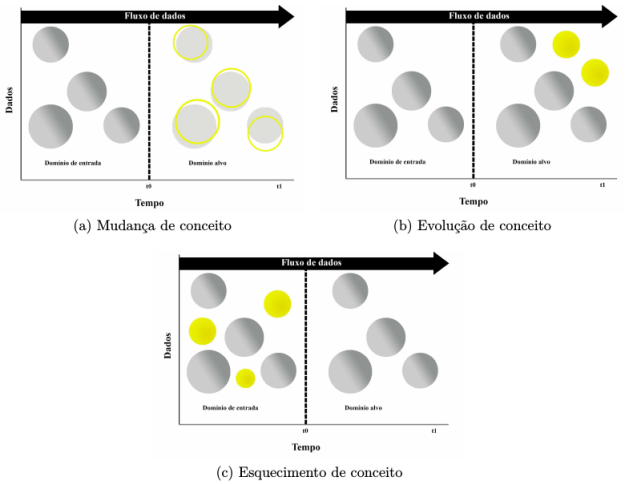
- **Sistemas de Segurança**: monitoramento contínuo através de imagens ou outros sensores para identificação de intrusos;
- **Redes de Computadores**: análise do tráfego de rede, monitorando pacotes distoantes, além de realizar a detecção de invasores;
- **Mercado Financeiro**: análise de dados e estatísticas da bolsa de valores, produzindo informações importantes para investidores. Outra vertente é a aplicação na detecção de fraudes;
- **Medicina**: aprimoramento do modelo de detecção de determinada doença a partir das análises e resultados de novos casos.

Conforme (GAMA et al., 2014), as principais características dos fluxos contínuos de dados são:

- **Contínuos**: Elementos que compõem os dados são recebidos de forma continuada;
- **Não estacionário**: Distribuição de probabilidade sofre alterações o longo do tempo;
- **Potencialmente Infinitos**: Os fluxos são potencionalmente infinitos, o que impede o completo armazenamento em memória.

Essas propriedades inviabilizam a aplicação dos algoritmos tradicionais de mineração de dados (MD) e aprendizagem de máquina, a fluxos de dados contínuos.

Mudança de conceito (*concept drift*) é uma mudança na distribuição dos dados utilizados para construção do modelo - definição dos conceitos (classes) - durante a execução da aplicação. A figura 1.1 representa um caso clássico de mudança de conceito: a alteração do perfil de compra do cliente ao longo do tempo.



(a) Mudança de conceito (b) Evolução de conceito

(c) Esquecimento de conceito

Figura 1.1 Mudança de conceito - Exemplo: Perfil de Compra

A evolução de conceito (*concept evolution*) caracteriza-se pela aparição de novas classes, diferente das classes conhecidas. Essa nova classe representa uma evolução, por exemplo, um novo interesse do cliente. Para que seja possível aprimorar modelos baseados em FCDs, é necessário esquecer conceitos desatualizados ou obsoletos, que apenas ocupam espaço e degradam o resultado das predições (ABDALLAH et al., 2016).

1.0.3 Algoritmos de classificação e FCDs

Algoritmos de classificação aplicados a FCDs permitem predizer, com alta precisão, a classe de novos exemplos obtidos a partir do fluxo. Durante o aprendizado supervisionado, um conjunto de dados previamente rotulado é fornecido ao algoritmo, para construção do modelo. A construção do modelo possibilita inferir a classe de novos exemplos que venham a ser encontrados. É importante que o modelo seja constantemente atualizado (evolua), para que contemple e classifique de forma correta novas distribuições de dados.

Dentre os algoritmos de classificação para cenários tradicionais (informação em lote, *batch*), estão: árvores de decisão, SVM e Naive Bayes. Esses algoritmos são aplicados em ambientes estacionários, isto é, em ambientes em que o modelo de decisão não requer atualizações e o algoritmo pode considerar que todos dados necessários podem ser armazenados em memória.

Entretanto, o aprendizado de novos conceitos a partir de fluxos contínuos de dados ocorre de forma signitificamente diferente do modelo tradicional, estático. Para lidar

REFERÊNCIAS BIBLIOGRÁFICAS

ABDALLAH, Z. et al. Anynovel: detection of novel concepts in evolving data streams. v. 7, p. 73–93, 06 2016.

AL-KHATEEB, T. M. et al. Cloud guided stream classification using class-based ensemble. In: *2012 IEEE Fifth International Conference on Cloud Computing*. [S.l.: s.n.], 2012. p. 694–701. ISSN 2159-6182.

BABCOCK, B. et al. Models and issues in data stream systems. In: *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York, NY, USA: ACM, 2002. (PODS '02), p. 1–16. ISBN 1-58113-507-6. Disponível em: ⟨http://doi.acm.org/10.1145/543613.543615⟩.

GAMA, J. a. et al. A survey on concept drift adaptation. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 46, n. 4, p. 44:1–44:37, mar. 2014. ISSN 0360-0300. Disponível em: ⟨http://doi.acm.org/10.1145/2523813⟩.

SPINOSA, E. J.; CARVALHO, A. P. de Leon F. de; GAMA, J. a. Novelty detection with application to data streams. *Intell. Data Anal.*, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 13, n. 3, p. 405–422, ago. 2009. ISSN 1088-467X. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=1551768.1551770⟩.