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, ensejando 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 estes 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 aumento da frequência e da quantidade dos dados obtidos originou a ideia de ambientes dinâmicos, completamente diferentes das bases de dados tradicionais, que supõem cenários estáticos. Os avanços tecnológicos também permitiram o surgimento de diversas aplicações do mundo real com essas características:

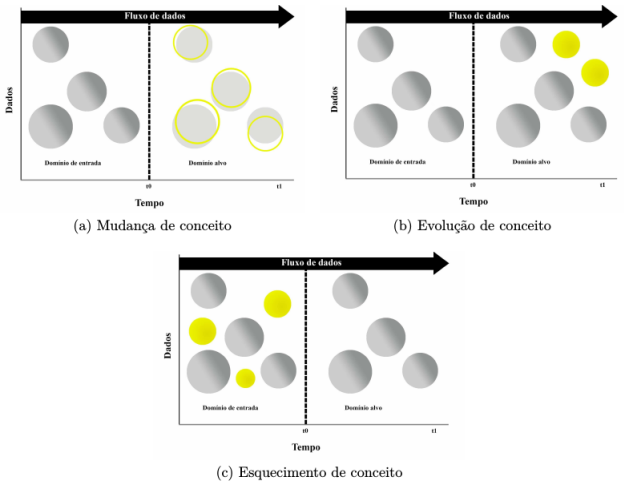
- **Sistemas de Segurança**: monitoramento contínuo através de imagens ou outros sensores para identificação de intrusos;
- **Redes de Computadores**: análise do tráfego de rede, monitorando pacotes distoantes, além de realizar a detecção de invasores;
- **Mercado Financeiro**: análise de dados e estatísticas 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 em cenários com FCDs.

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 significativamente diferente do modelo tradicional, estático. Para lidar com esse novo cenário, diversos algoritmos clássicos têm sido adaptados e novos algoritmos estão sendo desenvolvidos.

Algoritmos de aprendizagem de máquina com foco em FCDs geralmente atuam em duas fases: *offline* e *online*. Na fase *offline*, o algoritmo recebe exemplos rotulados, que são utilizados na elaboração de um modelo de decisão ou atualização do modelo existente. Na fase *online*, novos exemplos são classificados conforme são recebidos. Ao longo do tempo, o modelo construído é continuamente atualizado, para evitar que perca acurácia ou torne-se obsoleto.

A maior dificuldade para os algoritmos de classificação em cenários FCDs é a detecção e tratamento de mudanças de conceito (AGGARWAL, 2006), ou seja, mudanças na distribuição dos dados ao longo do tempo. Além desse, outros desafios precisam ser tratados, tais como: evolução de conceitos, detecção e remoção de *outliers* e ruídos, além da manutenção de uma boa precisão paralelo à evolução dos dados. Os algoritmos devem ter a capacidade de continuamente atualizar o modelo de decisão, permitindo à aplicação um processo de aprendizagem contínuo.

O processo de atualização do modelo de decisão é um componente importante dos algoritmos de classificação em FCDs. Este melhoramento pode ser feito com ou sem *feedback* externo. Algoritmos que utilizam *feedback* externo assumem que o rótulo verdadeiro de todos exemplos estará disponível, mesmo que seja com um certo atraso. Nesta abordagem, o algoritmo atualiza o modelo de tempos em tempos, conforme novos rótulos de exemplos são obtidos. Os algoritmos que não utilizam *feedback* em seu ciclo de atualização, não requerem a correta rotulação pós-processamento. Na abordagem sem *feedback*, outros indicadores e métodos são utilizados para renovação do modelo. Uma terceira abordagem

é a utilização parcial de *feedbacks*, . Neste caso, um conjunto parcial de rótulos corretos é utilizado para atualizar o modelo.

1.0.4 Detecção de Novidades em FCDs

Detecção de novidade (*concept drift detection*) consiste em identificar quando a distribuição dos dados de teste diverge da distribuição utilizada em treinamento. Esta é uma tarefa importante em classificação de FCDs, graças à natureza não-estacionária destes. A detecção de novidades tem como principal objetivo perceber novos padrões emergentes no fluxo de dados, possibilitando a identificação de novos conceitos, mudanças nos conceitos conhecidos e a presença de ruídos.

Na literatura recente, é possível encontrar algoritmos para classificação em cenários FCD com suporte à detecção de novidades (DN), entre os mais mencionados estão: (MASUD et al., 2010), (SPINOSA; CARVALHO; GAMA, 2009), (HAYAT; HASHEMI, 2010) e (FARID; RAHMAN, 2012). O desenvolvimento de algoritmos para DN em cenários FCD tem como principais desafios:

- **Contextos Recorrentes**: Conceitos ora esquecidos ressurgem, podendo ser confudidos com "novos"conceitos;
- **Mudança de Conceito**: Dificulta a correta classificação dos exemplos ao longo do tempo;
- **Evolução de Conceitos**: Novas classes surgem com o decorrer do tempo;
- **Ruídos ou *Outliers***: Pontos distoantes na distribuição que podem ser confudidos como novos conceitos.

Geralmente, a detecção de novidades (DN) ocorre durante a fase *online*, através da detecção de exemplos que não podem ser identificados pelo modelo atual. Entretanto, algumas propostas simplesmente consideram estes exemplos como atípicos. Outras propostas, mais refinadas, salvam estes exemplos em uma memória temporária e utilizam-os em uma análise futura, objetivando identificar o surgimento de uma novidade.

Apesar da maioria dos algoritmos para detecção de novidades em FCDs considerar que apenas uma classe nova pode surgir a cada janela de tempo considerada, alguns trabalhos como (MASUD et al., 2010), entendem que diferentes padrões novos podem surgir a cada intervalo, sendo importante diferenciá-los.

Conforme já explicitado, a atualização do modelo pode utilizar ou não *feedback* externo. Uma abordagem elegante, é a utilização de técnicas de aprendizado ativo para escolher, dentre os exemplos não classificados pelo modelo, quais devem ser rotulados e utilizados no processo de atualização.

1.0.5 Aprendizado Ativo e FCDs

A tarefa de classificação pode utilizar uma metodologia chamada de aprendizagem ativa (*active learning*), que possibilita ao algoritmo separar parte das instâncias e solicitar a

correta rotulação a um especialista (RICCARDI; HAKKANI-TUR, 2005). As instâncias rotuladas são então aplicadas no processo de atualização do modelo de decisão. A utilização da técnica de aprendizagem ativa busca minimizar o custo de rotulação dos dados para atualização do modelo. É um estratégia interessante, sobretudo em cenários FCDs, pois não requer a rotulação de todos os dados para manutenção do modelo.

Por não rotular todos os dados, o maior desafio da aprendizagem ativa é como selecionar as instâncias a serem rotuladas, para que se possa obter maior e melhor capacidade de previsão (RICCARDI; HAKKANI-TUR, 2005).

Na literatura é possível obter uma lista das capacidades que as estratégias de aprendizagem ativa devem apresentar (IENCO; PFAHRINGER; ZLIOBAIT, 2014):

- Preservar a distribuição dos dados de entrada;
- Detectar onde ocorrem as mudanças ao longo do fluxo de dados;
- Equilibrar, ao longo do tempo, o custo de realizar a rotulação das instâncias.

Novos algoritmos têm sido desenvolvidos para lidar com a escassez de rótulos para atualização de modelos que lidam com fluxos contínuos. O particionamento dos fluxos em lotes é a principal estratégia adotada por estas novas abordagens. Assim, o processo de aprendizagem ativa se dá por lote. Os dados presentes em cada lote são considerados estacionários. O objetivo, é treinar um classificador mais preciso a partir de uma pequena porção dos dados, diminuindo o custo para rotular.

Outra abordagem, também muito aplicada, é o agrupamento de exemplos não classificados pelo modelo, assumindo-os como potenciais novidades. Rotular, então, um representante de cada grupo e utilizá-lo para manutenção do modelo. Esta estratégia é aplicada, por exemplo, no algoritmo AnyNovel (ABDALLAH et al., 2016).

1.0.6 Algoritmos para DN em FCDs

Os algoritmos para detecção de novidades em fluxos de dados contínuos são divididos em duas fases: *offline* e *online*. Na fase *offline*, o modelo de decisão é construído a partir de dados previamente rotulados. Enquanto que na fase *online*, este modelo é aplicado para inferir os exemplos recebidos a partir do fluxo.

Nesta seção serão descritos alguns dos algoritmos para DN em FCDs presentes na literatura. Serão abordados os seguintes algoritmos: AnyNovel (ABDALLAH et al., 2016), ECSMiner (MASUD et al., 2010), CLAM (AL-KHATEEB et al., 2012) e OLINDDA (SPINOSA; CARVALHO; GAMA, 2009).

O AnyNovel (ABDALLAH et al., 2016), é um algoritmo para tratamento de *concept drift*, especializado em um cenário específico: a identificação de atividades em fluxos de dados gerados por sensores. O algoritmo faz uso da metodologia de aprendizagem ativa. Os dados recebidos são armazenados em uma memória temporária. Esta memória é dividida em blocos que são liberados para análise apenas quando o número máximo de instâncias é atingido. O algoritmo, então, classifica os dados do bloco entre: existente, novidade ou desconhecido. A aprendizagem ativa, através de um especialista, ocorre quando blocos de novidade ou desconhecidos são encontrados.

O ECSMiner (MASUD et al., 2010) realiza o tratamento de DNs em tarefas de classificação multiclasse. O ECSMiner constrói um comitê de classificadores de árvoes de decisão a partir de *chunks* dos dados. Os nós-folhas das árvores construídas são agrupados. Estes microgrupos são utilizados para explicar os exemplos do fluxo. Exemplos que não estejam enquadrados em nenhum subgrupo, são marcados como *outliers* e armazenados para futura análise. Quando um número suficiente de *outliers* é detectado, e estes *outliers* formam um grupo distinto dos grupos já conhecidos, ocorre a identificação de uma novidade. A principal limitação do algoritmo é a suposição de que no máximo apenas uma novidade ocorrerá por *chunk* de dados.

O algoritmo CLAM (AL-KHATEEB et al., 2012) atua de forma similar ao ECSMiner, mas ao invés de utilizar apenas um comitê, um comitê de classificadores é criado para cada classe conhecida do problema. Cada classificado de cada comitê é formado por um conjunto de microgrupos. A classificação de um exemplo é realizada através da identificação de qual microgrupo apresenta o centro mais próximo do exemplo. A cada exemplo, cada um dos comitês tenta classificar o exemplo, e caso não seja possível, marca-o como desconhecido. Para que um exemplo seja definitivamente marcado como desconhecido, ele tem que ser marcado assim pela maioria dos comitês. Se assim for classificado, resta identificado como um novo conceito. Assim como o ECSMiner, o algoritmo pressupõe que somente um novo padrão será identificado por bloco de dados.

O OLLINDDA (SPINOSA; CARVALHO; GAMA, 2009) é um algoritmo para DN que considera que na fase offline, somente exemplos da classe normal estão disponíveis para treinar o modelo de decisão inicial. Na fase *offline* o sub-modelo normal é criado. Os sub-modelos extensão e novidades são criados durante a fase *online*.

O OLLINDDA utiliza um algoritmo de agrupamento para produzir **k** grupos, representados por centro e raio. Este conjunto de **k** grupos representa o modelo normal. Um macro-esfera também é criada para manter um macrogrupo dos grupos criados. Esta macro-esfera realiza a separação do modelo normal dos sub-modelos de extensão e novidade. Quando novos exemplos são classificados dentro da macro-esfera, eles são considerados como extensão. Se forem classificados fora da macro-esfera, são percebidos como novidade. A classificação de novos exemplos se dá pelo cálculo da distância do centroide do grupo encontrado até o centroide do grupo mais próximo. Se a distância calculada for menor que o raio do grupo, este exemplo pode ser explicado pelo modelo atual, senão será marcado como desconhecido.

1.0.7 Considerações Finais

Neste capítulo foram apresentados alguns algoritmos para detecção de novidades em fluxos de dados contínuos identificados na literatura. Verificou-se que a maior parte dos algoritmos lida com mudanças de conceito através atualização do modelo ao longo da execução através de estratégias supervisionadas ou não supervisionadas.

Contudo, novos algoritmos utilizando a metodologia de aprendizagem ativa têm surgido. Nos trabalhos que apresentam tais algoritmos, percebe-se a obtenção de maior eficácia, considerando a quantidade de erros na classificação e a capacidade preditiva do algoritmo.

REFERÊNCIAS BIBLIOGRÁFICAS

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

AGGARWAL, C. C. *Data Streams: Models and Algorithms (Advances in Database Systems)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387287590.

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⟩.

FARID, D. M.; RAHMAN, C. M. Novel class detection in concept-drifting data stream mining employing decision tree. In: *2012 7th International Conference on Electrical and Computer Engineering*. [S.l.: s.n.], 2012. p. 630–633.

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⟩.

HAYAT, M. Z.; HASHEMI, M. R. A dct based approach for detecting novelty and concept drift in data streams. In: *2010 International Conference of Soft Computing and Pattern Recognition*. [S.l.: s.n.], 2010. p. 373–378.

IENCO, D.; PFAHRINGER, B.; ZLIOBAIT, I. High density-focused uncertainty sampling for active learning over evolving stream data. In: *Proceedings of the 3rd International Conference on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications - Volume 36*. JMLR.org, 2014. (BIGMINE14), p. 133–148. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2999973.2999984⟩.

MASUD, M. M. et al. Addressing concept-evolution in concept-drifting data streams. In: *Proceedings of the 2010 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2010. (ICDM '10), p. 929–934. ISBN 978-0-7695-4256-0. Disponível em: ⟨http://dx.doi.org/10.1109/ICDM.2010.160⟩.

RICCARDI, G.; HAKKANI-TUR, D. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 4, p. 504–511, July 2005. ISSN 1063-6676.

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⟩.