UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2023

EVOLUÇÃO DE CLASSIFICADORES FUZZY CONSIDERANDO FLUXO DE DADOS

Vanderleicio Carvalho Leite Junior[1]; Matheus Giovanni Pires [2]

1. Bolsista PIBIC/CNPq, Graduando em Engenharia da Computação, Universidade Estadual de Feira de Santana, e-mail: vanderleiciojr397@gmail.com

2. Orientador, Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: mgpires@ecomp.uefs.br

PALAVRAS-CHAVE: classificação; fluxo de dados; *concept drift*; evolução sistemas fuzzy.

INTRODUÇÃO

Com a crescente tecnológica que o mundo tem vivenciado, cada vez mais dados são gerados ao longo do tempo, e boa parte deles precisam ser computados e interpretados logo após sua geração, caracterizando os chamados fluxos de dados (Agrahari & Singh, 2021). Atrelada a isso, está a necessidade de adaptação dos modelos usados para classificar esses dados, tendo em vista que eles não têm um limite finito e por isso podem, e normalmente vão, variar ao longo do tempo (Frías-Blanco et al, 2014). A essa variação é dado o nome de *concept drift* (Korycki & Krawczyk, 2021) e sistemas inteligentes precisam se adaptar a essas mudanças.

Os Sistemas Baseados em Regras Fuzzy (SBRF) são amplamente usados em diversas áreas, e têm como principal ponto a favor a sua capacidade de modelar o raciocínio de forma aproximada, através do uso da Lógica Fuzzy (Mendel, 1995). Um SBRF aplicado na resolução de problemas de classificação são chamados Sistemas de Classificação Baseados em Regras Fuzzy (SCBRF). Um dos principais aspectos na construção de SBRF é a definição da base de regras, pois parte do conhecimento do sistema é representado pelas regras.

Considerando o contexto de fluxo de dados, as variações das informações ao longo do tempo e a necessidade de manter os sistemas sempre aptos na classificação dos dados atuais, este trabalho tem por objetivo propor formas de ajustar a base de regras de um SBRF na tarefa de classificação de fluxo de dados, mesmo diante de mudanças abruptas ou graduais desses dados.

METODOLOGIA

Como neste trabalho foi considerado dados não estacionários, inicialmente foi necessário escolher um algoritmo capaz de detectar possíveis *concepts drifts*. Após pesquisa na literatura sobre algoritmos de detecção de *concepts drifts*, escolhemos o algoritmo *Hoeffding Drift Detection Method* (HDDM) (Frías-Blanco et al, 2014), um método estatístico de detecção de *concept drift* que utiliza a desigualdade de Hoeffding. Este

algoritmo analisa uma das métricas de classificação do modelo para determinar se houve ou não mudança de conceito, com base na variação dela. As duas abordagens desse algoritmo foram usadas, sendo a HDDMA para a detecção de desvios de conceitos abruptos (usou-se o valor de 0,005 para o parâmetro "*confidence for the drift*", valor determinado a partir de testes), e a HDDMW para os graduais (com o valor de 0,01 para o parâmetro "*confidence for the drift*", também determinado a partir de testes). A escolha foi feita com base no tipo de desvio para o qual cada abordagem se adapta melhor e a sua utilização foi através do framework MOA, *Massive Online Analysis* (BIFET et al, 2010).

Os datasets utilizados neste trabalho foram os mesmos usados no trabalho de FRÍAS-BLANCO et al. (2015), com exceção do dataset "poker" que não está presente nesse artigo, mas que também é amplamente usado. Além disso, também foram gerados datasets através do framework MOA. Portanto, no total foram utilizados 15 datasets, dos quais três deles são reais (covtype, electricity e poker). O tipo de geração dos datasets variou entre abrupto, com 100 exemplos entre os conceitos, abrupto com 1000 exemplos entre os conceitos, gradual com 100 exemplos entre os conceitos e gradual com 1000 exemplos entre os conceitos. Vale ressaltar, que embora tenham 100 e 1000 exemplos entre cada mudança de conceito, os datasets graduais começam a mudar a partir de 50 e 500 exemplos antes, respectivamente, assim como feito por FRÍAS-BLANCO et al. (2014). Para a simulação da continuidade de chegada de dados, as instâncias foram divididas em janelas, assim para a janela zero de cada dataset foram selecionados os 50 primeiros dados, para a janela um os 50 dados seguintes, e assim por diante, até que todos os dados tivessem sido agrupados em janelas. Então, o SCBRF é criado com a janela zero e é usado para classificar os dados das janelas seguintes.

O algoritmo de Wang-Mendel (Wang e Mendel 1992) foi usado para gerar as regras do classificador fuzzy na primeira janela de cada execução e para classificar os dados das janelas seguintes. Três cenários foram propostos com o objetivo de tentar melhorar o classificador após a detecção de um desvio de conceito, são eles: "Cenário 1", a base de regras é recriada usando somente os dados da janela atual; "Cenário 2", regras com utilidade menor do que a utilidade média da base de regras são descartadas, e em seguida, a base é incrementada com novas regras; e o "Cenário 3", a base é incrementada com as novas regras e só depois a sua utilidade média é calculada e as regras com utilidade menor do que a média são descartadas.

Para avaliar o desempenho do modelo foram testadas quatro métricas, acurácia, G-mean, F1-Score e a área sobre a curva ROC (AUC). A necessidade do uso dessas métricas foi avaliada frente à presença de datasets com dados desbalanceados e com mais de duas classes, sendo assim, F1-Score foi a métrica escolhida para realizar a comparação entre os cenários.

RESULTADOS E/OU DISCUSSÃO

Nas Tabelas 1 e 2 são mostrados os resultados do F1-Score médio de cada um dos cenários supracitados. De acordo com os resultados apresentados, é possível perceber que o Cenário 3, na maioria dos casos desempenha melhor do que os outros cenários, a exemplo do dataset Agrawal. Aplicando o algoritmo de Wilcoxon (Wilcoxon 1945) nos resultados do dataset Agrawal para 100 exemplos por conceito e com mudanças graduais (Tabela 1), obtivemos p-valor de 2,8311E-07 entre os Cenários 3 e 2, indicando uma

diferença estatística significativa, ou seja, o desempenho do Cenário 3 realmente é melhor que o Cenário 2. Por outro lado, comparando os resultados entre os Cenários 1 e 3 na mesma situação, obtivemos um p-valor de 0,156367855, sendo assim, não há diferença significativa entre os resultados dos Cenários 1 e 3.

No entanto, para o dataset Stagger, tanto para as mudanças de conceito graduais quanto para as abruptas, os melhores resultados foram obtidos pelo Cenário 1, no qual a base de regras fuzzy é inteiramente reiniciada após a detecção de um desvio. Uma possível razão para este comportamento é que o dataset Stagger possui apenas três atributos. Considerando que foi adotada neste trabalho uma modelagem com três conjuntos fuzzy para cada atributo de todos os datasets, no caso do Stagger, a quantidade máxima de regras possível são 27. Portanto, manter todas as regras possíveis na base se mostrou mais vantajoso do que excluir algumas delas.

Tabela 1 – F1-Score médio dos cenários de *Concept Drift* Graduais.

Algoritmo de detecção	Cenário	100 exemplos por conceito			1000 exemplos por conceito		
		Led	Stagger	Agrawal	Led	Stagger	Agrawal
HDDMW	1	**0,51216**	**0,80221**	0,59667	0,44182	**0,83626**	0,41307
	2	0,39812	0,71502	0,27069	0,35562	0,68461	0,33670
	3	0,48197	0,70100	**0,63185**	**0,53333**	0,81087	**0,62266**

Tabela 2 – F1-Score médio dos cenários de *Concept Drift* Abruptos.

Algoritmo de detecção	Cenário	100 exemplos por conceito			1000 exemplos por conceito		
		Led	Stagger	Agrawal	Led	Stagger	Agrawal
HDDMA	1	0,50488	**0,63819**	0,24583	0,45459	**0,84701**	0,27590
	2	0,50072	0,53938	**0,58212**	0,35935	0,68253	0,36804
	3	**0,50554**	0,62365	0,43852	**0,54584**	0,71135	**0,36810**

A Tabela 3 traz a comparação entre os cenários em datasets reais. Esse tipo de dataset apresenta um nível maior de dificuldade no problema de detecção de drift, principalmente porque não há a informação exata em que pontos do dataset o drift ocorre, ou mesmo se ele ocorre (GONÇALVES, 2013). Porém, nos resultados também é possível perceber que o Cenário 3 conseguiu melhorar o desempenho do classificador fuzzy, principalmente quando comparado aos outros dois cenários, com exceção do dataset Poker, no qual o Cenário 3, desempenha de forma semelhante aos demais.

Tabela 3 – F1-Score médio do modelo nos Datasets Reais.

Algoritmo de detecção	Cenário	Nome do dataset		
		Poker	Covtype	Electricity
HDDMW	1	**0,38937**	0,40559	0,02276
	2	0,34972	0,31274	0,03517
	3	0,34509	**0,51820**	**0,35937**
HDDMA	1	**0,38813**	0,40689	0,01357
	2	0,35730	0,30267	0,04709
	3	0,37103	**0,56488**	**0,29447**

CONSIDERAÇÕES FINAIS

Neste trabalho foi investigado três formas de ajustar um Sistema de Classificação Baseado em Regras Fuzzy aplicado na classificação de fluxo de dados. Mais precisamente, três formas diferentes de atualizar a base de regras foram avaliadas. A primeira consistiu de recriar toda a base de regras usando os dados da janela atual, quando a mudança de conceito foi detectada. Na segunda forma, na detecção da mudança do conceito, uma medida de utilidade era calculada para cada regra da base de regras atual. As regras com valor de utilidade menor que a média da base de regras eram excluídas. Em seguida, novas regras eram adicionadas. Por fim, a terceira estratégia é semelhante à segunda. Na detecção da mudança do conceito, a base é incrementada com as novas regras e só depois a sua utilidade média era calculada e as regras com utilidade menor do que a média eram descartadas.

Os resultados obtidos puderam nortear uma possível alternativa de pesquisa mais aprofundada no que diz respeito ao ajuste de SCBRF na classificação de fluxo de dados. Foi possível perceber que dentre as estratégias propostas e diante dos cenários de drift e dos dados disponíveis, a ideia de implementar uma atualização da base de regras fuzzy através da comparação da utilidade das regras com a utilidade média da base, mostra-se promissora, quando comparada às outras duas estratégias implementadas.

No entanto, em outras situações, estratégias diferentes podem aprimorar melhor o modelo de classificação, como foi constado para o dataset Stagger. Neste caso, inferimos que em datasets mais simples, ou seja, com poucos atributos, não seja necessária a implementação de uma técnica mais robusta de renovação da base, e que somente refazê-la pode ser suficiente.

REFERÊNCIAS

- AGRAHARI, S.; SINGH, A. K. Concept Drift Detection in Data Stream Mining: A literature review. **Journal of King Saud University - Computer and Information Sciences**, 2021.
- BIFET, A. et al. MOA: Massive Online Analysis. **Journal of Machine Learning Research**, v.11, n.52, p. 1601–1604, 2010.
- FRIAS-BLANCO, I. et al. Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds. **IEEE Transactions on Knowledge and Data Engineering**, v.27, n.3, p.810–823, 2015.
- GONÇALVES JÚNIOR, P. M. Multivariate non-parametric statistical tests to reuse classifiers in recurring concept drifting environments**. Tese de Doutorado, Universidade Federal de Pernambuco**, disponível em https://repositorio.ufpe.br/handle/123456789/12226, acesso em 06/09/2023.
- KORYCKI, Ł.; KRAWCZYK, B**. Concept Drift Detection from Multi-Class Imbalanced Data Streams**, disponível em https://arxiv.org/abs/2104.10228.
- MENDEL, J. M. Fuzzy Logic Systems for Engineering: A Tutorial. **IEEE Special Issue on Fuzzy Logic**, vol.83, n.3, pp.345-377, 1995
- P. Angelov and Xiaowei Zhou, On line learning fuzzy rule-based system structure from data streams, **IEEE International Conference on Fuzzy Systems**, pp.915-922, 2008.
- Wang, L.x., Mendel, J.M.: Generating Fuzzy Rules by Learning from Examples. **IEEE Transactions on Fuzzy Systems, Man, And Cybernetics**, vol.22, n.6, pp.1414-1427, 1992.
- Wilcoxon, F. Individual comparisons by ranking methods. **Biometrics Bull**., vol.1, n.6, pp.80-83, 1945.