ALGORITMO DE APRENDIZAGEM DE MÁQUINA NÃO SUPERVISIONADO PARA DETECÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS MULTIVARIADAS APLICADAS A TEMPORADAS DE FURACÕES

Ilan Sousa Figueirêdo[1]; Lílian L. Nani Guarieiro[2]; Alex Álisson Bandeira Santos[2]; Erick Giovani S. Nascimento[2]
[1] Bolsista; PD&I2; ilan.figuegmailiredo@hotmail.com
[2] Professor; Centro Universitário SENAI CIMATEC; Salvador-BA; erick.sperandio@fieb.org.br

RESUMO

O objetivo do trabalho foi apresentar um método para detecção de anomalias em séries temporais multivariadas utilizando aprendizagem de máquina não supervisionada. O método é baseado em técnicas de agrupamento usando a função de distância Mahalanobis. Para avaliar sua eficácia, o método foi aplicado a um caso real correspondente a uma temporada de furacões no oceano Atlântico. O período analisado abrange os três principais eventos daquela temporada que passaram perto da localização da coleta de dados, os furacões Isaac, Rafael e Sandy. Os resultados obtidos sugerem que esta proposta pode ser aplicada com sucesso para detectar anomalias em séries temporais multivariadas.

PALAVRAS-CHAVE: detecção de anomalias; séries temporais; multivariado; aprendizagem de máquina não supervisionado

1. INTRODUÇÃO

Atualmente, muitos processos como plantas industriais, estações de monitoramento meteorológico ou bolsas de valores, geram dados relevantes de séries temporais continuamente, porém a análise manual de dados multivariados em larga escala é inviável para o homem, por conseguinte, surge a necessidade de automatizar o processo de análise para estruturar os dados e direcionar os especialistas para uma tomada de decisão mais confiável.[1]

Em muitos cenários, é essencial para o processo identificar padrões incomuns que podem ser gerados por comportamento imprevisível, o comportamento indesejado pode ser devido a algum problema que possa estar ocorrendo no processo relacionado. Por exemplo, em ambiente industrial as empresas podem usar os dados de monitoramento do estado de saúde de ativos a fim de identificar os comportamentos de operações anormais. Outro exemplo seria uma operadora de cartão de crédito que pode monitorar cada transação do usuário para procurar comportamentos incomuns que possam apontar para operações fraudulentas. Esses comportamentos indesejados e anormais são frequentemente chamados de comportamentos anômalos e podem ser extraídos nos dados devido a uma diversidade de razões, todas apresentando um certo grau de relevância para o analista. É importante que essa análise leve em consideração quaisquer alterações no comportamento do parâmetro para identificar oportunidades para melhorar, prevenir ou corrigir qualquer situação.[2]

A meteorologia é a área da ciência que estuda processos atmosféricos para fins de simulação e de previsão do tempo através dos seus fenômenos físico-químicos. Os fenômenos meteorológicos estão relacionados com múltiplas variáveis atmosféricas, como por exemplo: temperatura, pressão atmosférica, umidade do ar, velocidade e direção do vento, formação de nuvens, precipitação pluviométrica radiação de ondas curtas e longas, dentre outras, assim como suas relações e variações com o passar do tempo.[3] Entretanto, reconhecer comportamentos anômalos a partir de um grande conjunto de dados coletados e desempenhar previsões e classificações do tempo, torna-se inviável a um especialista sem auxílio de ferramentas adequadas. A título de exemplo, a detecção iminente de furacões que se aproximam da costa terrestre é fundamental para garantir a segurança pública. Por conseguinte, surge-se a necessidade de tecnologias rápidas e assertivas para proporcionarem um maior entendimento dos eventos meteorológicos, assim como suas variáveis para uma tomada de decisão.

Os dados meteorológicos podem ser caracterizados como multivariados, o que possibilita, por um lado, uma representação mais robusta dos fenômenos envolvidos. Por outro lado, apresenta um desafiado maior para aplicação de modelos de aprendizagem de máquina capazes de reconhecer padrões e prever o comportamento, devido ao maior volume de dados e atributos a serem correlacionados.

Neste contexto, o objetivo deste estudo foi detectar anomalias em séries temporais multivariadas, o que é uma tarefa essencial em diversas áreas de estudo, por exemplo, em estudos climáticos e ecossistêmicos,[4] pesquisa oceânica[5] ou em processos industriais.[6] Foi utilizado o algoritmo Cluster-based Algorithm for Anomaly Detection in Time Serie Using Mahalanobis Distante (C-AMDATS) como ferramenta para detectar furacões em dados de meteorologia.

O algoritmo C-AMDATS foi introduzido por Nascimento em 2015 como uma ferramenta de alto potencial para detecção de anomalias em séries temporais.[2] O algoritmo é baseado em aprendizagem de máquina não supervisionado e possui a distância de Mahalanobis como sua função de distância.

2. METODOLOGIA

O C-AMDATS é um algoritmo de agrupamento de aprendizagem de máquina não supervisionado, a técnica apresenta apenas dois hiperparâmetros: (i) tamanho inicial de agrupamento (TIA) e (ii) fator de agrupamento (FA). Nesse sentido, o TIA agrupa as sequências observadas dos dados, onde cada agrupamento pode representar um status de comportamento da série temporal. Após o agrupamento inicial, é reconstruído um melhor agrupamento no conjunto de dados conforme a distribuição de pontos ao longo do tempo, esse fenômeno de melhor reagrupamento dos dados é devido à distância de Mahalanobis incorporada no algoritmo. O uso da função de distância Euclidiana em algoritmos de agrupamento é comum, a sua função exerce com que os agrupamentos assumam a forma geométrica de um círculo, pois não leva em consideração a variação de cada dimensão do conjunto de dados. No entanto, há situações em que o agrupamento em círculo não representa bem os dados (por exemplo: quando há grande variação nos eixos X e Y nos dados). Portanto, a distância de Mahalanobis leva em consideração as variações de cada dimensão. A ferramenta tem potencial para auxiliar o especialista na tomada de decisões em diversas áreas da engenharia e saúde. A Equação (1) apresenta a formula da distância de Mahalanobis.[2]

$$d_m(x,\mu) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \qquad (1)$$

Na Equação (1), $x = (x_1, x_2, ..., x_n)^T$ é uma variável específica no dataset, onde n é o número de dimensão das variáveis, $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$, é um certo centroide de agrupamento e S é matriz de covariância relativa a esse agrupamento.

Depois que os agrupamentos foram encontrados por completo na série temporal, a etapa a seguir realiza a atividade de encontrar os padrões ocultos *P* na série temporal *T*. O algoritmo examina mutuamente a similaridade dos agrupamentos, essa similaridade é calculada pelo produto entre o desvio padrão σ_y dos valores reais das variáveis em *T* (coordenada-y) de cada agrupamento com o FA. Portanto, se o módulo da diferença entre a coordenada-y dos centroides de dois agrupamentos for menor ou igual a similaridade, esses agrupamentos poderão ser mesclados, o que significa que representarão o mesmo padrão *P*. Essa atividade é executada até todos os agrupamentos serem analisados.

Na terceira e última etapa, o algoritmo realiza a detecção de anomalias. Uma anomalia é um padrão que não está em conformidade com o comportamento conjecturado em *T*, *i.e*, um padrão anômalo. Essa detecção é realizada através do cálculo de uma pontuação que mede a anomalia *r* para cada padrão *P* (encontrado na etapa anterior). A pontuação é calculada entre a razão do tamanho de toda a série temporal pelo somatório dos tamanhos dos agrupamentos presentes em *P*. A pontuação de anomalia *r* avalia o grau de relevância de *P* no que tange a detecção de anomalia. Então, todo o conjunto *P* é ordenado por *r* em ordem decrescente, e os padrões anômalos serão aqueles com os maiores valores de pontuação de anomalia. Quanto maior o valor da pontuação de anomalia para um padrão *P*, maior é a probabilidade de ser uma anomalia em *T*.

Este trabalho implementou o método C-AMDATS na linguagem de programação Python 3.6 e o executou no ambiente de computação de alto desempenho Ogum, localizado no Centro de Supercomputação para Inovação Industrial do SENAI CIMATEC. O modelo de processador é uma CPU Intel (R) Xeon (Gold) 6148 Gold a 2,40GHz e possui 187 GB de memória RAM.

Para avaliação de desempenho do algoritmo, o experimento foi realizado em um caso real, assim sendo, este trabalho utilizou o dataset público Meteocean data da National Data Buoy Center pertencente da National Oceanic and Atmospheric Administration's[1] (NOAA). Os dados foram coletados no oceano Atlântico nas proximidades da costa de Bahamas (23.838 N, 68.333 O). O dataset possui aproximadamente seis meses de dados e estão estruturados na frequência horária, iniciando em junho de 2012 até novembro de 2012 (213 dias e 22 horas), totalizando 15.315 pontos de dados. Este período corresponde à temporada de furacões no Atlântico, que naquele ano foi especialmente ativa com 19 ciclones tropicais (ventos acima de 52 km/h), dos quais 10 se tornaram furacões (ventos acima de 64 km/h). Por serem conjunto de dados reais, as anomalias afetam diversas variáveis ao mesmo tempo, essas foram: (a) altura significativa da onda, (b) pressão do nível do mar e (c) velocidade do vento.

3. RESULTADOS E DISCUSSÃO

Executamos o método C-AMDATS nas três variáveis supracitadas e comparamos os resultados com os furacões históricos nas Bahamas. As caixas coloridas tracejadas representam a duração oficial dos três principais eventos daquela temporada que passaram perto da localização, os furacões Isaac, Rafael e Sandy,

[1] http://www.ndbc.noaa.gov/

respectivamente. A Figura 1 ilustra os padrões detectados pelo modelo. Observe que, em geral, as áreas verdadeiras do histórico são maiores que as detecções, pois abrangem a vida inteira dos furacões.

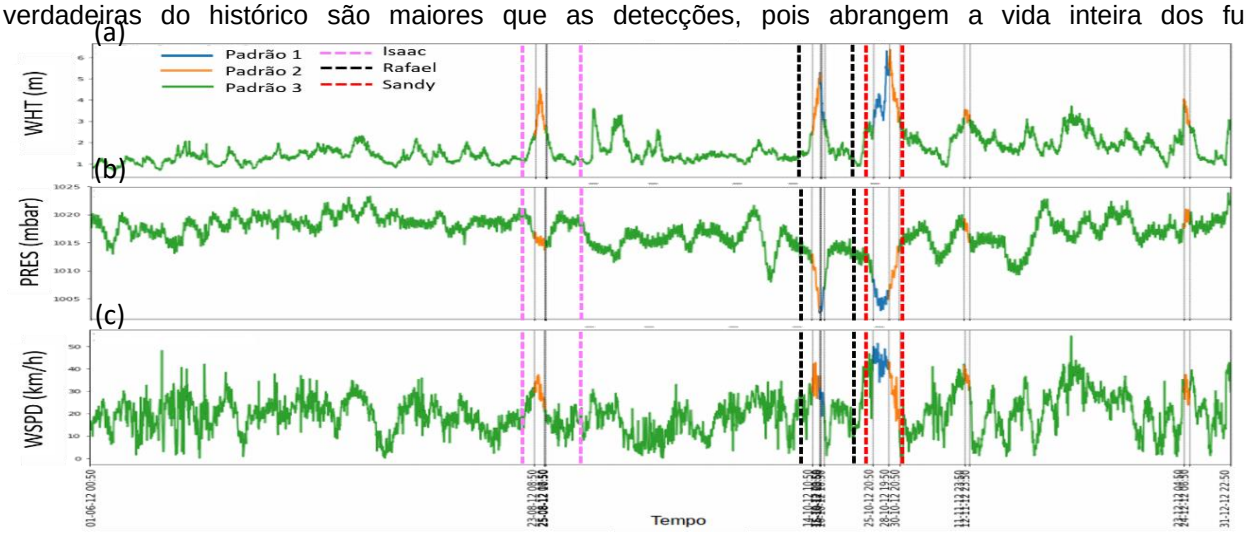

Figura 1. Visualização da detecção de padrões pelo C-AMDATS na série temporal. As linhas verticais sólidas representam o intervalo de tempo dos padrões detectados. (a) altura significativa da onda, (b) pressão do nível do mar e (c) velocidade do vento. As caixas tracejadas em cores representam os históricos dos furacões Isaac, Rafael e Sandy.

O C-AMDATS identificou três padrões de interesse na série temporal multivariada, precisamente nos períodos de vida dos furacões Isaac, Rafael e Sandy. Os falsos positivos após Sandy podem estar relacionados tanto a uma tempestade local ou às reminiscências de outras anomalias. Os valores dos parâmetros utilizados para FA e TIC foram 3,0 e 24, respectivamente.

Para avaliar o desempenho do C-AMDATS em relação a capacidade de identificar os mesmos padrões anômalos já identificados pelos especialistas humanos, foram utilizados sete metodologias de avaliação, os resultados foram: (i) Accuracy 90%, (ii) Precision 81%, (iii) Recall 33%, (iv) Specificity 98%, (v) F1-score 74%, AUCROC 66% e AUCPRC 62%. Não foi identificado outro estudo que realizou uma avalição quantitativa usando técnicas de aprendizagem de máquina não supervisionado para detecção de anomalias nesse conjunto de dados.

4. CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um método para detectar anomalias em séries temporais. Apresentamos os conceitos por trás do algoritmo e em seguida descrevemos nossa abordagem. Por fim, executamos o algoritmo para um caso real e identificamos bons resultados no desempenho. O método permite a detecção eficiente de padrões de anomalias em séries temporais multivariadas e apresentou ser uma ferramenta útil para a descoberta de dados na ciência natural.

Agradecimentos

Agradecemos a NOAA que fornece ao público dados meteorológicos para pesquisa. Agradecemos a Empresa Brasileira de Pesquisa e Inovação Industrial (EMBRAPII) pelo financiamento parcial desta pesquisa.

5. REFERÊNCIAS

1. Rodner, E. *et al.* Maximally Divergent Intervals for Anomaly Detection. in *ICML Workshop on Anomaly Detection* (2016). doi:10.17871/BACI_ICML2016_Rodner
2. Nascimento, E. G. S., De Lira Tavares, O. & De Souza, A. F. A cluster-based algorithm for anomaly detection in time series using mahalanobis distance. *Proc. 2015 Int. Conf. Artif. Intell. ICAI 2015 - WORLDCOMP 2015* 622–628 (2015).
3. Vieira, N. R. *Poluição do ar: indicadores ambientais*. (Editora E-papers, 2009).
4. Zscheischler, J. *et al.* Extreme events in gross primary production: a characterization across continents. *Biogeosciences* **11**, 2909–2924 (2014).
5. Mínguez, R., Reguero, B. G., Luceño, A. & Méndez, F. J. Regression Models for Outlier Identification (Hurricanes and Typhoons) in Wave Hindcast Databases. *J. Atmos. Ocean. Technol.* **29**, 267–285 (2012).
6. Darkow, T., Dittmar, R. & Timm, H. Real-time application of multivariate statistical methods for early event detection in an industrial slurry stripper. *IFAC Proc. Vol.* **47**, 8879–8884 (2014).