Identificação de Falhas em Turbinas Eólicas Utilizando Abordagens de Aprendizado de Máquina

Danielle R. Pinna[1], Rodrigo F. Toso[3], Kele Belloze [1], Fernando de Sá[2], Raphael Guerra[2], Diego N. Brandão[1],

[1]Programa de Pós-graduação em Ciência da Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ
Rio de Janeiro – RJ – Brasil

[2]Instituto de Computação - UFF, Niterói

[3]Microsoft, USA

danielle.pinna@eic.cefet-rj.br, diego.brandao@cefet-rj.br

Abstract. *The last few years have been marked by the insertion of renewable technologies in the global energy matrix, such as wind and solar energy, considered clean energies with low environmental impact. Wind turbines, responsible for the energy conversion process, are complex, high-cost equipment and susceptible to numerous failures. Monitoring turbine components can help detect failures before they occur, reducing equipment maintenance costs. This work compares data-centric machine-learning techniques in fault detection in wind turbines. Results show the importance of data selection and optimization in the problem context.*

Resumo. *Os últimos anos têm sido marcados pela inserção de tecnologias renováveis na matriz energética mundial, como a energia eólica e solar, que são energias limpas e de baixo impacto ambiental. As turbinas eólicas, responsáveis pelo processo de conversão energética, se constituem por equipamentos complexos de alto custo e suscetíveis a inúmeras falhas. O monitoramento dos componentes das turbinas pode auxiliar na detecção de falhas antes que elas ocorram, reduzindo os custos de manutenção do equipamento. Este trabalho compara duas técnicas para determinação de hiperparâmetros de modelos centrados em dados na detecção de falhas em turbinas eólicas. Resultados mostram a importância da seleção e otimização de dados para o problema.*

1. Introdução

A energia eólica é um recurso de energia renovável disponível na natureza que vem sendo cada vez mais utilizada, principalmente por ser uma energia limpa. No Brasil, essa fonte de energia terminou o ano de 2022 com um crescimento de 18,8% em relação ao ano anterior[ABEEólica 2022]. A turbina eólica é responsável pela transformação da energia eólica em energia elétrica. Entretanto, a operação e manutenção das turbinas representam cerca de 25% a 35% dos custos de geração [Blanco et al. 2017].

Os problemas relacionados à manutenção da turbina eólica normalmente envolvem falhas do sistema elétrico e as provenientes das condições climáticas extremas. A falha do componente ocasiona redução da produtividade ou até mesmo o desligamento da

turbina. Por essa razão, a maneira mais eficaz de reduzir os custos de manutenção é monitorar o *status* dos geradores e prever o seu mau funcionamento antes que o sistema falhe [Qin et al. 2017]. Assim, o diagnóstico precoce de falhas é um fator chave para reduzir significativamente os custos de manutenção.

Atualmente, os aerogeradores modernos já possuem um sistema de coleta e de armazenamento de dados, conhecido como sistema de controle e aquisição de dados - SCADA, que monitora e armazena dados de todo o funcionamento das turbinas por meio de sensores instalados em seus componentes. A vantagem em lidar com a medição de dados em tempo real é que eles representam o estado de saúde real da turbina, que está diretamente relacionado à possibilidade de redução dos custos de manutenção.

A maioria dos estudos sobre detecção de falhas em turbinas eólicas utiliza conjuntos de dados operacionais e de eventos, como os fornecidos pelo SCADA [Stetco et al. 2019]. Esses dados podem ser usados para o desenvolvimento de modelos de detecção de falhas em aerogeradores, como os modelos de Aprendizado de Máquina (AM). Um modelo de AM usa dados de treinamento para aprender a relação entre dados de entrada e saída, e pode ser usado para classificar novos dados de entrada. Uma das partes mais importantes do AM é selecionar os hiperparâmetros ideais para garantir um modelo preciso e eficiente.

Este trabalho visa comparar duas técnicas de ajuste de hiperparâmetros em modelos de AM, considerando uma abordagem centrada em dados, para auxiliar na detecção de falhas dos componentes das turbinas eólicas extraídas a partir do SCADA. A novidade do manuscrito é uma aplicação de métodos de otimização de hiperparâmetros de AM em um problema real. O artigo está organizado em mais 4 seções. A seção 2 apresenta os trabalhos relacionados. A seção 3 descreve a metodologia. Os resultados são discutidos na seção 4 e, por fim, a seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

Nos últimos anos, com o crescente volume de dados gerados e com uma maior complexidade dos problemas a serem tratados computacionalmente, tornou-se necessário o uso de ferramentas computacionais mais sofisticadas e autônomas. Um exemplo são os algoritmos de AM que são capazes de modelar os dados e resolver problemas complexos.

Uma revisão da literatura sobre os modelos de AM usados no monitoramento de turbinas eólicas, incluindo a tarefa de detecção de falhas pode ser vista em [Stetco et al. 2019]. Os autores analisaram as etapas de AM: fonte de dados, seleção de atributos, escolha e validação do modelo e a tomada de decisão. Eles concluíram que a maior parte dos trabalhos usa dados oriundos do SCADA com modelos de classificação.

Em [Garan et al. 2022], os autores enfatizam que mais esforços deveriam ser colocados na qualidade do conjunto de dados a fim de melhorar o desempenho das medidas de classificação. Além disso, eles propõem uma metodologia centrada em dados, onde as etapas orientadas a dados são executadas de forma iterativa na detecção de falhas das componentes das turbinas eólicas.

Uma revisão sobre o estado da arte de dados SCADA para o monitoramento de turbinas eólicas é feita em [Pandit et al. 2023]. Os autores destacam que a seleção de recursos pode aumentar a precisão do modelo e reduzir o tempo computacional, visto que

os dados SCADA contêm muitas variáveis redundantes.

Como pode ser observado, os trabalhos da literatura apresentados não abordam sobre a determinação dos hiperparâmetros dos modelos desenvolvidos. Assim, este trabalho realiza uma introdução neste assunto, comparando duas técnicas de determinação de hiperparâmetros no contexto de identificação de falhas em turbinas eólicas.

3. Metodologia

A Figura 1 ilustra a metodologia adotada: coleta e análise da base de dados, pré-processamento dos dados, seleção de atributos, treinamento do algoritmo de AM supervisionado junto com o ajuste dos hiperparâmetros e a avaliação de desempenho do modelo.

Figura 1. *Pipeline* da metodologia adotada.

A base de dados é fornecida pela empresa Energias de Portugal (EDP) [EDP 2021]. Esse é um dos conjuntos de dados gratuitos mais completos para análise de recursos eólicos e pesquisa do desempenho de turbinas eólicas [Mendes et al. 2020]. Os registros consistem em informações extraídas do SCADA, de cinco turbinas eólicas medidos nos anos de 2016 e 2017, contendo: (i) *Signals*: *Dataset* das variáveis dos sensores do sistema SCADA medido a cada 10 minutos; (ii) *Metmast*: *Dataset* das variáveis do mastro meteorológico; (iii) *Failures*: *Dataset* com o registro das ocorrências de falhas de cinco componentes da turbina eólica.

A etapa inicial de pré-processamento envolve a inclusão dos dados meteorológicos na base de dados dos sensores pelo tempo de medição. Em alguns instantes de tempo, a não aquisição de dados pelos sensores por algum motivo exigiu o uso de técnicas de imputação de dados, para que a série estivesse completa com todas as medições a cada 10 minutos. Após essa etapa, incluiu-se os dados das falhas pelo código da turbina e tempo de medição menor ou igual ao tempo de falha.

Este trabalho realiza a classificação do estado de saúde da turbina eólica, em que o rótulo de classe igual a '1' é atribuído ao conjunto de dados coletado 60 dias antes da ocorrência da falha e o rótulo de classe igual a '0' é atribuído para os conjuntos de dados coletados nos demais intervalos relativos ao registro da falha. O limite de 60 dias foi definido a partir de uma avaliação divulgada pela EDP, que determinou o período de 60 dias precedentes à falha como razoável para identificar nos dados capturados o comportamento que indica a falha iminente.

No pré-processamento, ainda foi necessário que os atributos com pouca variação e denominados como *offset* fossem removidos do conjunto de dados. Além disso, com o intuito de remover dados redundantes, apenas os valores médios de cada variável foram selecionados, totalizando 60 *atributos*. A normalização dos dados numéricos foi feita para eliminar a discrepância das unidades de medida entre as variáveis.

Nas técnicas de seleção ou extração de atributos comparamos dois métodos, a Informação Mútua (MI) e a Análise de Componentes Principais (PCA). A MI mede a quantidade de informação que uma variável dá sobre a outra, ou seja, mede a dependência entre as variáveis. O PCA é usado para eliminar a alta correlação e reduzir a dimensionalidade dos dados multivariados com perda mínima de informação.

Para a última etapa, utilizou-se os classificadores supervisionados: Regressão Logística (RL), K-vizinhos mais próximos (kNN), Árvore de Decisão (AD), Floresta Aleatória (FA) e Máquinas de Vetores de Suporte (SVM), que são alguns dos algoritmos mais usados na tarefa de classificação para aprender com dados de turbinas eólicas [Pandit et al. 2023]. Um método bastante utilizado para analisar os resultados produzidos pelos classificadores é a matriz de confusão e as medidas de desempenho que dela resultam, como: Acurácia, Revocação, Precisão, F_1score, AUC (Área sob a Curva ROC) e Coeficiente de Matthews. Todas essas medidas são usadas para medir a qualidade das classificações.

A otimização de hiperparâmetros, ou ajuste, é o processo de encontrar a combinação certa de valores de hiperparâmetros para obter o máximo desempenho de dados em um período de tempo razoável. Um pré-requisito para treinar modelos de AM em geral é criar uma combinação específica de valores de hiperparâmetros. Somente após a escolha de um conjunto específico de hiperparâmetros é que o processo de treinamento pode ajustar os parâmetros do modelo [Japa et al. 2023]. Isso pode ser particularmente importante ao comparar o desempenho de diferentes modelos de AM em um conjunto de dados.

Na busca pela melhor combinação dos hiperparâmetros, serão avaliadas duas técnicas de otimização para medir o tempo computacional de treinamento dos modelos. O *Random Search* é uma abordagem popular, que testa combinações aleatórias de um intervalo dos hiperparâmetros. Este método pode não encontrar o melhor conjunto de hiperparâmetros, mas pode fornecer um modelo que se aproxima do ideal em termos de desempenho, economizando muito tempo computacional. Uma desvantagem da pesquisa aleatória é que ela não tenta melhorar com base em combinações de hiperparâmetros previamente testadas [Japa et al. 2023].

O *Halving Random Search* implementa uma estratégia de torneio de forma sucessiva. Isto quer dizer que ele começa com um pequeno número de casos de treinamento para identificar e selecionar rapidamente modelos candidatos pouco promissores. Os modelos que sobrevivem para a próxima rodada são avaliados usando uma proporção maior dos dados disponíveis. Esse processo se repete até restar apenas alguns modelos candidatos, que são então treinados e avaliados usando todos os dados disponíveis [Soper 2023].

A Figura 2 apresenta uma representação gráfica do algoritmo de *Halving Random Search* aplicado de forma sucessiva, no qual o número de modelos candidatos diminui exponencialmente de uma iteração para a próxima, enquanto o número de casos de treinamento aumenta exponencialmente de uma iteração para a próxima.

4. Resultados

Os experimentos foram realizados com rotinas computacionais implementadas em Python versão 3, em uma máquina Intel(R) Xeon(R) Gold 5120 CPU 2.20GHz, com 28 núcleos

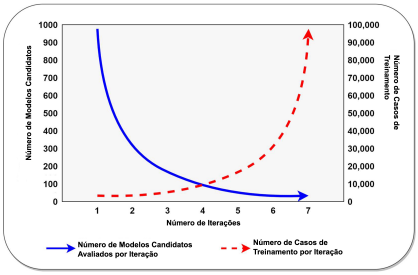

Figura 2. Representação do *Halving Random Search* sucessivo [Soper 2023].

e $192GB$ de memória. As bibliotecas utilizadas foram *pandas*[1], *numpy*[2] e *scikit-learn*[3]. A base de dados é particionada em 80% para treinamento dos modelos e 20% para teste, mantendo a ordem do conjunto de dados. Para encontrar o modelo mais otimizado, os métodos *Random Search* (RS) e *Halving Random Search* (HRS) com validação cruzada são utilizados para definir os hiperparâmetros.

A Tabela 1 apresenta as métricas de desempenho do melhor classificador na base de teste para cada otimizador com melhor técnica de seleção de atributos. Ela apresenta também o tempo computacional de treinamento e a combinação de valores dos hiperparâmetros de cada modelo com o maior F_1 *score*. Em todos os modelos das componentes da turbina, o método *halving* obteve um baixo custo computacional, com praticamente todas as métricas de desempenho ficando semelhantes ao *Random Search*. Para o componente Caixa de Velocidade, o uso do *halving* reduziu 4 horas de processamento.

Tabela 1. Resultado para cada otimizador e seleção de atributos na base de teste.

Componente	Modelo	Otimização de Hiperparâmetro	Seleção de Atributos	F_1 *score*	AUC	Acurácia	Precisão	Revocação	Coef. Matthews	Tempo Computacional	Hiperparâmetro
Caixa de Velocidade	RL	HRS	PCA	32,2%	57,6%	69,4%	62,3%	21,7%	22,2%	0 hrs 8 min 41 sec	penalty='l1', solver='liblinear'
	RL	RS	PCA	32,2%	57,5%	69,4%	62,2%	21,7%	22,2%	4 hrs 37 min 58 sec	penalty=None, solver='newton-cholesky'
Gerador	AD	HRS	MI	31,1%	56,4%	71,1%	45,7%	23,5%	16,4%	0 hrs 0 min 1 sec	criterion='entropy', max_depth=100, max_features='log2'
	AD	RS	MI	39,3%	60,6%	74,1%	55,8%	30,4%	26,5%	0 hrs 1 min 10 sec	max_depth=10, max_features='sqrt'
Rolamento do Gerador	RL	HRS	MI	16,7%	52,7%	66,8%	52,5%	9,9%	10,5%	0 hrs 0 min 26 sec	penalty='l1', solver='liblinear'
	RL	RS	MI	18,7%	53,3%	67,2%	55,0%	11,3%	12,4%	0 hrs 10 min 46 sec	penalty=None, solver='newton-cholesky'
Grupo Hidráulico	RL	HRS	PCA	17,5%	54,7%	62,0%	95,9%	9,6%	22,9%	0 hrs 0 min 17 sec	penalty='l1', solver='liblinear'
	RL	RS	PCA	16,9%	54,5%	61,9%	95,7%	9,3%	22,5%	0 hrs 20 min 20 sec	penalty='l1', solver='saga'
Transformador	RL	HRS	PCA	23,9%	56,5%	73,9%	85,6%	13,9%	27,6%	0 hrs 0 min 7 sec	solver='sag'
	RL	RS	PCA	23,9%	56,5%	73,9%	85,5%	13,9%	27,6%	1 hr 31 min 37 sec	penalty=None

A comparação dos resultados pela métrica do F_1 *score* do melhor modelo de cada componente da turbina eólica será feita com os resultados do artigo *benchmark* [Garan et al. 2022]. Naquele artigo os autores também utilizam uma abordagem centrada em dados com diferentes técnicas de seleção de atributos para cada componente com um modelo de Árvore de Decisão. Para os componentes da turbina eólica Transformador e Gerador, a abordagem aqui apresentada obteve resultados superiores aos do estudo de caso, enquanto que para os demais componentes o F_1 *score* foi menor, conforme pode ser observado na Tabela 2.

[1]https://pandas.pydata.org/
[2]https://numpy.org/
[3]https://scikit-learn.org/stable/

Tabela 2. Comparação da métrica F_1 *score* dos resultados para o *benchmark*.

Componente	*Benchmark*	Resultado
Caixa de Velocidade	37,7%	32,2%
Gerador	9,6%	**39,3%**
Rolamento do Gerador	36,3%	18,7%
Grupo Hidráulico	44,9%	16,9%
Transformador	8,1%	**23,9%**

5. Considerações Finais

A manutenção preditiva de máquinas que se desgastam com o tempo é um método importante para melhorar a eficiência do processo. As turbinas eólicas são sistemas complexos que demandam manutenção, e à medida que a aquisição de dados aumenta, também amplia a possibilidade de aplicação de algoritmos de aprendizado de máquina combinando abordagens centrada em dados, para melhorar a qualidade dos dados que entram nos modelos e otimizar o tempo de treinamento.

A detecção de falhas foi realizada em um conjunto de dados reais do sistema SCADA durante o monitoramento de turbinas eólicas. A seleção de atributos foi aplicada criando um subconjunto de variáveis com redução de dimensionalidade e perda mínima de informação. A metodologia proposta foi capaz de predizer as falhas dos componentes, e dois destes componentes superaram os resultados da literatura. Para trabalhos futuros podemos testar outros tipos algoritmos, como os da classe *ensembles* e também técnicas de sobreamostragem que considerem a dependência temporal dos dados.

Referências

ABEEólica (2022). Abeeólica - Associação Brasileira de Energia Eólica. `https://abeeolica.org.br/`, last accessed on 01/05/22.

Blanco, M. A. et al. (2017). Impact of target variable distribution type over the regression analysis in wind turbine data. *IWOBI 2017 - Proceedings*.

EDP (2021). EDP - Open Data. `https://opendata.edp.com/pages/homepage/`, last accessed on 15/08/21.

Garan, M. et al. (2022). A data-centric machine learning methodology: Application on predictive maintenance of wind turbines. *Energies*, 15:826.

Japa, L. et al. (2023). A population-based hybrid approach for hyperparameter optimization of neural networks. *IEEE Access*, 11:50752–50768.

Mendes, M. et al. (2020). Wind farm and resource datasets: A comprehensive survey and overview. *Energies*, 13.

Pandit, R. et al. (2023). Scada data for wind turbine data-driven condition/performance monitoring: A review on state-of-art, challenges and future trends. *Wind Engineering*, 47(2):422–441.

Qin, S. et al. (2017). Ensemble learning-based wind turbine fault prediction method with adaptive feature selection. *Comm. in Computer and Information Science*, 728.

Soper, D. S. (2023). Hyperparameter optimization using successive halving with greedy cross validation. *Algorithms*, 16(1).

Stetco, A. et al. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable energy*, 133:620–635.