Diagnóstico de falhas em caldeira de recuperação Kraft

Autores: Felipe Barbuda Gradin[1*]; Gustavo Ryuji Taira[2]; Henrique Camargo de Araújo Venturelli[c]; Song Won Park[1]

[1] *Departamento de Engenharia Química, Escola Politécnica, Universidade de São Paulo*

[2] *Petrobras, Rio de Janeiro, Brasil*

[3] *Unichem Processos, São Caetano do Sul, Brasil*

Autor Correspondente: [*]felipe.gradin@usp.br

Resumo

O presente trabalho tem como objetivo analisar o uso de métodos de Machine Learning como ferramenta para detecção de falhas em processos químicos. A indústria química possui uma série de regulações ambientais e de segurança, as quais tornam o gerenciamento dos processos químicos uma tarefa complexa. Por conseguinte, torna-se extremamente importante o desenvolvimento de um sistema de monitoramento de processo integrado e diagnóstico de anormalidades. Diante dessa questão, surge a possibilidade do uso de modelos preditivos a partir da análise de dados, que servem como uma alternativa de baixo custo para o diagnóstico dos processos.

Neste trabalho, foi utilizado o algoritmo Support Vector Clustering (Ben-Hur et al., 2001), baseado nas Support Vector Machines, para analisar os dados de uma caldeira de recuperação Kraft. Os dados da caldeira foram medidos em situações normais de funcionamento, e em situações com formação de incrustações. Depois, foram utilizados para treinar o algoritmo. Os testes apontam um bom funcionamento do programa, com eficiência maior que 90% nas predições. Foram utilizadas como medidas de desempenho cross-validation, F1 score e AUC, e foi feita uma comparação do programa com os algoritmos de classificação Naïve Bayes e Support Vector Classifier.

Já houve diversas instâncias de uso de SVM para diagnóstico de falhas em processos industriais (Yin e Hou, 2016), e este trabalho mostra que é viável a utilização de Support Vector Clustering como alternativa para dados não supervisionados. Uma limitação do estudo é a análise da proximidade da falha em que um dado se encontra, que é uma possibilidade com esse tipo de algoritmo e seria uma adição crucial para, além de detectar falhas rapidamente, poder preveni-las.

Palavras-chave

Support Vector Machine, Support Vector Clustering, Machine Learning, Fault Detection, Kraft Recovery Boiler

1. Introdução

Em geral, os trabalhos de modelagem de processos são feitos com base em modelos matemáticos que representam as equações fenomenológicas em engenharia química, como balanços de massa e energia, cinética de reação e assim por diante. Ao longo de modelos de processos satisfatórios, diversas pesquisas incluem otimizações de processos. Se a modelagem fenomenológica não está disponível devido à falta do conhecimento do processo, ou porque a sua elaboração é custosa com alta incerteza, sempre se pode elaborar modelos substitutos, desde que se tenha uma grande quantidade de dados com qualidade (Almeida et al., 2022). No presente caso da operação de uma caldeira de recuperação química Kraft, é necessário criar diretamente a detecção de suas falhas, uma vez que o uso de modelos, fenomenológicos ou substitutos, conduz a métodos não eficientes de análise de residuais. Entende-se aqui como situação de falha aquela em que ocorre operação não desejada pelos operadores na fábrica, e não exclusivamente uma situação que exige parada de equipamento. O principal interesse é demonstrar a aplicação de *Support Vector Clustering* (Ben Hur et al; 2001), comparado com técnicas clássicas de SVM, *Support Vector Machines* (Yin e Hou; 2015; Glauner et al., 2017) e Naive Bayes (Wu e Kumar, 2009).

2. Caldeira de Recuperação Química Kraft

2.1. Funcionamento da Caldeira

A caldeira de recuperação Kraft é formada por uma fornalha, superfícies de convecção (superaquecedor, economizador e balões), e equipamentos auxiliares, como precipitador eletrostático e tanque de mistura. As principais atribuições da Caldeira de Recuperação Química (Vakkilainen, 2005) são: combustão da

fração orgânica do licor negro para reduzir poluentes, geração de vapor (possibilitando tanto o uso de vapor vivo de processo quanto a cogeração de energia elétrica em turbinas), produção de sulfeto de sódio a partir da redução de compostos inorgânicos de enxofre, gerando fundidos (*smelt*), e obtenção de licor verde, a partir do processamento da fração inorgânica (smelt) para recuperação de soda como licor branco fraco. Uma caldeira de recuperação de licor preto possui três regiões importantes: leito de carbonizado com fundidos de inorgânicos onde se tem a zona de redução, fornalha ou zona de combustão do licor, e parte superior da caldeira, região dos acessórios (superaquecedores, balões e economizadores) ou região de transferência de calor convectiva. Enquanto uma caldeira para gerar vapor e realizar co-geração de eletricidade, as duas últimas são mais importantes. A Figura 1 mostra um esquema representativo de uma caldeira de recuperação moderna ilustrando as principais regiões, o nariz da caldeira e a localização das entradas de ar e dos injetores de licor. A estrutura denominada nariz da caldeira, delimita a fronteira entre a fornalha e a região dos acessórios (Ribeiro et al., 2007, Ferreira et al, 2009). Nesta Figura 1 ilustra-se tipicamente a localização dos níveis de alimentação de ar para uma caldeira moderna.

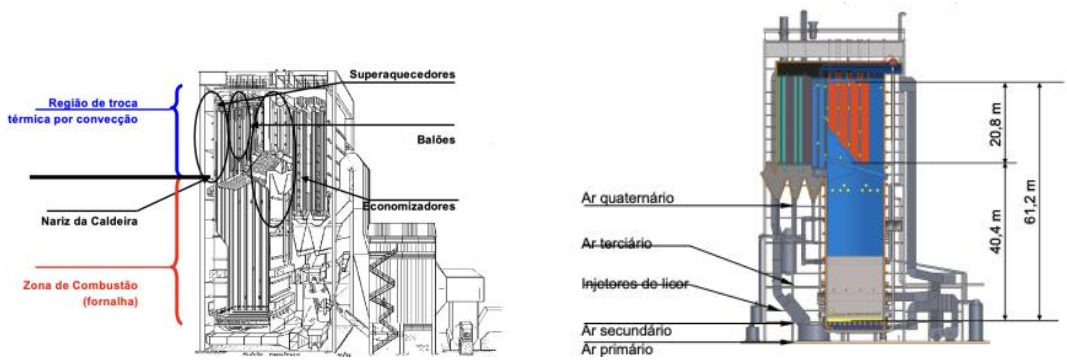

Figura 1: Representação esquemática de uma caldeira de recuperação química. Fonte: (Ribeiro *et al*. 2007, Ferreira et al. 2009).

2.2. Coleta de dados

Os dados utilizados são provenientes de um levantamento feito em 2020 na caldeira de recuperação Kraft. Foram feitas medições em intervalos de um minuto ao longo de dois dias: no primeiro, o cenário da caldeira era estável, e no segundo, verificou-se a formação de incrustações, totalizando 5760 instantes de tempo de operação (2880 por dia). Os dados apresentavam 88 variáveis do processo da caldeira, e foram marcadas 18 como as mais relevantes quanto à ocorrência ou não de incrustações. A Tabela 1 ilustra estas variáveis escolhidas a partir da experiência da operação. A incrustação é algo que chega a exigir parada de manutenção em casos críticos, porque a caldeira é um meio intensivo em particulados que podem até sinterizar nos bancos de tubos. A partir desses dados, foi feito um conjunto de treinamento, composto igualmente por casos de falha e funcionamento normal. Depois, foi separado um conjunto de testes, com casos escolhidos aleatoriamente para avaliar a performance do algoritmo em sua tarefa de classificação.

Tabela 1: Variáveis que definem a incrustação na caldeira

TAG	719FI870	719FI874	719FI878	719TI141	719TI138	719TI884
	Vazão ar primário	Vazão ar secundário	Vazão ar terciário	Temperatura água após econ/entrada balão	Temperatura vapor saída 2 estágio superaquecedor/ entrada dessuper	Temperatura gases entrada dos evaporadores - lado norte
	t/h	t/h	t/h	C	C	C
8/12/19 0:00	113,0645523	155,6368408	101,0911407	228,8026733	401,79776	380,0519714
8/12/19 0:01	112,9985352	155,6077728	100,7472076	228,8026733	401,79776	379,4382019
8/12/19 0:02	113,5181427	153,8927155	101,6950302	228,4162598	402,5596924	379,4382019

TAG	719TI885	719PI884	719PI885	719TI882A	719TI882B	719PI883	
	Temperatura gases entrada dos evaporadores - lado sul	Pressão gases saída evaporadores	Pressão gases saída evaporadores	Temperatura gases saída dos superaque-cedores - lado esquerdo	Temperatura gases saída dos superaque-cedores - lado direito	Pressão gases saída superaque-cedores	
	C	mmH2O	mmH2O	C	C	mmH2O	
8/12/19 0:00	386,072113	-11,21371365	-17,14185143	468,2584534	400,3595581	-58,79915619	
8/12/19 0:01	387,0224915	-11,36937714	-16,9276638	468,2584534	399,382019	-58,79915619	
8/12/19 0:02	387,0224915	-11,03347111	-17,7890892	468,2584534	399,382019	-58,79915619	
TAG	**719PI882**	**719MVTC004_I**	**719PCV881**	**719PCV882**	**719PI881**	**719FI879**	
	pressão gases na região dos superaquecedores	Corrente motor do ventilador de tiragem PE/F	Abertura damper IDF PE/F	Abertura damper IDF PG	pressão fornalha	gas tq dissolvedor p/ ar secundário	**Incrustação**
	mmH2O	A	%	%	mmH2O	t/h	
8/12/19 0:00	7,275283337	160,7173462	25	63,49671173	-0,1130559295	19,95389175	1
8/12/19 0:01	8,79096508	160,7930756	25	63,39321136	-0,1762580872	21,4566803	1
8/12/19 0:02	9,341060638	158,4398499	25	62,44224167	1,385304809	23,68124008	1

3. Métodos

3.1. Naïve-Bayes

O método Naïve-Bayes é um algoritmo conhecido por ser fácil de construir, não necessitar de esquemas complicados de estimativa de parâmetros iterativos, ser aplicável para bases de dados de grande extensão, ser fácil de se interpretar e, ainda mais importante, apresentar bom desempenho. Ele pode não ser o melhor classificador para todas as aplicações, mas pode ser normalmente considerado um modelo confiável, robusto e com boa precisão. Omite-se a descrição por fórmulas aqui, por concisão das páginas, e recomenda-se ver em Wu e Kumar (2009). A ideia base do algoritmo Naïve Bayes é considerar que uma estimativa de probabilidade bayesiana $P(i|x)$, por si própria, seria um valor adequado para ser utilizado como regra de classificação. O núcleo do método Naïve Bayes está no método de estimar o $P(i|x)$. O método de Naïve Bayes assume que os componentes de x são independentes dentro de cada classe, sendo daí a origem do nome alternativo de "*independence* Bayes". Uma questão que chama muita a atenção do algoritmo Naïve Bayes é o fato de se assumir independência dos atributos dentro de uma classe e esta condição não ocorrer na maioria dos casos reais. A priori, seria esperado que, devido à suposição desta hipótese improvável, o método tivesse uma performance fraca. Entretanto, o que normalmente ocorre é que o modelo consegue realizar classificações relativamente precisas na maioria dos casos reais em que é aplicado. A razão pela qual a suposição de independência não é tão prejudicial quanto parece é que fatores relacionados à natureza do modelo e ao estado dos dados utilizados tornam esta suposição pouco relevante nos resultados do algoritmo.

3.2. Support Vector Machines

As *Support Vector Machines*, ou SVM, são algoritmos que utilizam vetores de suporte para tarefas de classificação e regressão, e estão entre os mais robustos e precisos métodos aplicados para mineração de dados. Compostas por classificadores de vetores suporte (*Support Vector Classifier* – SVC) ou regressores de vetores suporte (*Support Vector Regressor* – SVR), as SVM foram originalmente desenvolvidas por Vapnik em meados de 1990 com base teórica fundada na teoria de aprendizado estatístico, exigindo apenas um número reduzido de amostras para treinamento e apresentando grande eficiência em amostras com número elevado de dimensões. Aqui também será seguido o texto de Wu e Kumar (2009), omitindo-se muitas passagens.

3.2.1. Support Vector Classifier

Para uma base de dados com duas classes linearmente separáveis, o objetivo do *Support Vector Classifier* é achar um hiperplano capaz de separar as duas classes das amostras com uma margem máxima que possa

oferecer a melhor capacidade de generalização. Capacidade de generalização se refere à característica de um classificador não apenas ter um bom desempenho de classificação dos dados de treinamento, mas também garantir uma elevada precisão preditiva para dados futuros. Intuitivamente, a margem pode ser definida como a quantidade de espaço, ou separação, entre as duas classes definida por um hiperplano. Geometricamente, a margem corresponde a menor distância entre os pontos mais próximos a qualquer ponto do hiperplano.

3.2.2. Support Vector Clustering

Support Vector Clustering, também baseado nas *Support Vector Machines*, busca formar um limite de decisão (com a inclusão de uma margem para outliers) para a clusterização de dados multidimensionais. A diferença para o SVM é que em vez de utilizar um hiperplano, o SVC mapeia os pontos para um *feature space* de dimensão alta, onde procura a menor esfera que encapsule os dados. Para isso, é utilizado um kernel gaussiano, com parâmetro de largura q (para o SVC da caldeira, foi utilizado $q = 0,25$). Para diferenciar pontos de acordo com seus clusters, é feita uma matriz de adjacência, em que todos os pontos dentro de um raio menor que o da esfera recebem o valor de 1, e, fora desse raio, 0.

A ideia por trás do uso de um algoritmo de clustering neste projeto é aplicar métodos para dados não supervisionados, e extrapolar diferentes condições de operação do sistema para determinar a proximidade de falhas operacionais. Como os dados recebidos possuem classificação binária, decidiu-se testar se haveria a formação de dois clusters a partir deles. Isso ocorreu para uma parte dos dados, mas em outras porções houve formação de mais clusters, indicando que há diferenças mais sutis na operação, ou que existe a necessidade de ajustar os parâmetros do algoritmo.

4. Resultados

Diversas medidas de eficiência podem ser aplicadas para avaliar o desempenho de um modelo. Algumas delas incluem precisão: $\frac{VP}{(VP + FP)}$, revocação: $\frac{FP}{(VP + FN)}$ e F1-score: $\frac{2 \times \text{precisão} \times \text{revocação}}{\text{precisão} + \text{revocação}}$, sendo que VP = verdadeiro positivo, FP = falso positivo e FN = falso negativo. O F1-score é a média harmônica das duas medidas, e serve como uma medida unificada para avaliar os resultados. Como o algoritmo trata do diagnóstico de falhas, deve-se dar mais importância à revocação, já que seria mais prejudicial classificar uma operação anormal da planta como normal do que o contrário. Tais medidas foram incluídas dentro do código do SVC, e cada uma teve desempenho superior a 90%.

A fim de testar a eficiência do Support Vector Clustering em comparação a outros algoritmos, foram testados dois algoritmos diferentes com o mesmo conjunto de treinamento, sendo estes o *Naïve-Bayes Gaussiano* e o *Support Vector Classifier*, ambos retirados da biblioteca SciKit-Learn. Como foi dito anteriormente, o *Support Vector Clustering* não manteve a formação de apenas dois clusters para todos os dados, então a presente análise foi feita considerando os primeiros 100 casos de falha e os primeiros 100 casos de operação normal.

O método *Naïve-Bayes* apresentou uma alta taxa de revocação (em torno de 90%) e uma baixa precisão (em torno de 50%), ou seja, foi eficiente em diagnosticar falhas, mas acusou anomalias em muitos casos onde não existiam. O modelo SVM teve resultados similares. Mostrou-se perfeito no diagnóstico de falhas (revocação de 100%), mas acusou casos de anomalias que não existiram, obtendo precisão em torno de 70%.

Segue abaixo um exemplo de matriz de confusão para o *Naive-Bayes* e para o *Support Vector Classifier*.

Tabela 2: Matriz de Confusão

	Naïve Bayes		SVC	
	Predição SEM incrustação	**Predição COM incrustação**	**Predição SEM incrustação**	**Predição COM incrustação**
Real SEM incrustação	2%	47%	26%	22%
Real COM incrustação	10%	41%	0%	52%

Nenhum dos dois algoritmos teve o mesmo problema que o *Support Vector Clustering* ao adicionar mais dados ao conjunto de treinamento, e suas precisões aumentaram significativamente com o aumento de dados de treinamento.

5. Conclusão

O algoritmo escolhido se mostrou eficiente na tarefa de classificação apresentada, porém há uma possibilidade de que tenha ocorrido sobreajuste dos dados, já que não houve um volume grande de dados analisados. Além disso, o funcionamento do *Support Vector Classifier*. e do *Naïve-Bayes* para um conjunto maior de dados indica que o *Support Vector Clustering* precisa ter seus parâmetros ajustados, ou que há diferenças mais sutis nas operações da caldeira do que a classificação binária de falhas exibe.

Em relação a sua aplicabilidade, existe outro ponto importante: o projeto foi feito com um conjunto de dados em batch, ou seja, os dados foram tratados offline. Sistemas de aprendizado em batch têm certas limitações, pois eles não aprendem de forma incremental, ou seja, devem ser treinados usando todos os dados disponíveis, depois destes terem sido coletados. Para acessar dados novos, então, o sistema deve ser descontinuado e atualizado para que a análise possa acontecer. Já que o modelo em questão deve diagnosticar falhas no processo em tempo real, isso pode-se mostrar trabalhoso.

A vantagem do uso de *Support Vector Clustering* é que, por ser um método de aprendizado não supervisionado (ou seja, não depende de classe previamente definidas para diagnóstico de falhas), é mais adequado para acompanhar PHD (*Process Historical Data*).

O presente trabalho não apresentou nenhum gráfico temporal porque, com 18 variáveis, um gráfico de uma variável por vez não reflete a situação da operação. Fica como a sequência, o desafio de apresentar as condições de falha em hiperespaço.

Agradecimentos

Os autores agradecem ao CNPq pelos recursos de Processo 127195/2023-8.

Referências

de Almeida, G. M., Park, S. W., & Lee, C. J. (2022). A Stochastic Optimization based on Sample Average Approximation for a Boiler Process. The 13th Dycops Dynamics and control of process systems. Busan, Jun, 14-17, 2022. Busan. Korea *IFAC-PapersOnLine, 55*(7), 550-555.

de Almeida, G. M., Park, S. W., Kim, S. C., & Lee, C. J. (2020). A sampling-based stochastic optimization for a boiler process in a pulp industry. *Korean Journal of Chemical Engineering*, 37, 588-596.

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). *Support vector clustering. Journal of machine learning research, 2*(Dec), 125-137.

Ferreira, D. J. D. O., Cardoso, M., & Park, S. W. (2010). Gas flow analysis in a Kraft recovery boiler. *Fuel Processing Technology, 91*(7), 789-798.

P. Glauner, M. Du, V. Paraschiv, A. Boytsov, I. Lopez Andrade, J. Meira, P. Valtchev, and R. State. The top 10 topics in machine learning revisited: A quantitative meta-study. In *Proceedings of the 25th European*

Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017), pages 299–304, 2017.

Ribeiro, R. N., Muniz, E. S., Lamarque, L. H. F., Mehta, R. K., & Park, S. W. (2007). Automação e sistemas de segurança em caldeiras de recuperação química. *Intech Brasil BLRBAC, no95*, 7-22.

Vakkilainen, E. *Kraft Recovery Boilers–Principles and Practice*. Lutpub Lappeenranta University of Technology, Finland. 2005

Wu, X., & Kumar, V. (eds.). (2009). *The top ten algorithms in data mining*. CRC press.

Yin, Z., & Hou, J. (2016). Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. *Neurocomputing, 174*, 643-650.