**PREDIÇÃO ONLINE DO TEOR DE ÁLCOOL NA VINHAÇA VIA SENSOR VIRTUAL BASEADO EM** *k-Nearest Neighbors***: UM ESTUDO DE CASO REAL**

**Paolla Marlene Caetano da Cunha¹; Gustavo Matheus de Almeida²**

**Resumo**

O monitoramento contínuo de variáveis-chave em processos industriais é crucial para garantir produção com qualidade e segurança operacional. Ainda, a construção de modelos através de uma descrição puramente matemática não é geralmente satisfatória, dado a complexidade das operações. Assim, a disponibilidade de dados e um conjunto de técnicas baseadas em dados constituem-se uma solução alternativa para a construção de sensores virtuais. Esse contexto também se refere as usinas de cana-de-açúcar, cuja automação tem gerado um volume significativo de dados. Na operação de destilação do bioetanol, o controle do teor alcóolico na vinhaça é importante por representar a maior perda alcóolico no processo. O infortúnio desta é ser medida em laboratório e, portanto, não ser disponibilizada imediatamente à operação. O conhecimento prévio desse valor, de forma contínua, agregaria a correção antecipada de operações anormais. O objetivo deste trabalho é a construção de um sensor virtual para o teor de álcool na vinhaça, a partir, exclusivamente, de variáveis com medição *online*, como vazão, pressão e temperatura. Com isso, obtém-se uma estimativa desse teor alcóolico a cada novo instante de amostragem dessas variáveis de processo. O estudo de caso refere-se a uma usina sucroenergética brasileira, cujo produto final é o etanol hidratado. Ao apresentar o melhor desempenho entre um conjunto de técnicas adotou-se o algoritmo supervisionado, k-vizinhos mais próximos (*k-Nearest Neighbors*; k-NN), para a construção do sensor virtual. O problema é composto por duas classes, valores aceitáveis e não aceitáveis do teor alcóolico. Após as etapas de pré-processamento dos dados, identificação e seleção de modelos, testou-se o modelo k-NN final em um conjunto de dados independentes, obtendo-se uma acurácia de aproximadamente 89%. Concluindo, a disponibilização *online* do teor alcóolico é essencial para se alcançar decisões mais rápidas e assertivas do ponto de vista operacional, contribuindo para uma maior eficiência em geral.

PALAVRAS-CHAVE: Destilação; Análise de dados; *k-Nearest Neighbors* (*k*-NN); Etanol hidratado.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Aluna de Mestrado do Programa de Pós-Graduação em Engenharia Química – Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, paollacaetano@gmail.com

2. Doutor – Grupo de Análise e Visualização de Dados, Departamento de Engenharia Química, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

**Introdução**

O uso de sensores para a medição de variáveis de campo em geral proporciona um monitoramento contínuo do processo. Nos últimos anos, o aumento significativo desses instrumentos nas indústrias vem gerando uma quantidade massiva de dados. Atualmente, a sua coleta e armazenamento já não são mais um problema; porém, como transformá-los em informação relevante para a organização, gerando conhecimento para tomadas de decisões mais racionais (ALMEIDA; PARK, 2017).

Os processos químicos industriais apresentam características como, não linearidade, alta quantidade de variáveis, interação entre componentes químicos, dependência com o tempo, vida útil de equipamentos, entre outras características. Nesse contexto, a modelagem fenomenológica se torna complexa e cara (em relação ao tempo e conhecimento necessários) para a descrição das operações, requerendo simplificações ou suposições de idealidade (LIMA et al., 2016).

Essas restrições impulsionaram o desenvolvimento e uso de modelos obtidos diretamente a partir de conjuntos de dados em diversas áreas, incluindo a engenharia de processos industriais. De modo resumido, a ciência de dados pode ser definida como uma área multidisciplinar, que busca transformar dados em conhecimento relevante (GILBERT et al., 2006).

Um exemplo de uso de técnicas de análise de dados em engenharia de processos diz respeito à construção de sensores virtuais. Essa demanda ocorre em várias situações. Por exemplo, quando não há um sensor físico online disponível, ou o seu custo é impeditivo, ou em caso de ser necessário estimar parâmetros de laboratório de modo contínuo. Lembra-se que é comum ter um atraso entre a coleta de amostra e a disponibilização do resultado de sua análise em laboratório para o time de processo. Esse atraso geralmente causa perdas, como por exemplo, produção fora da especificação.

O teor alcoólico no produto final é um parâmetro-chave no processo de destilação alcóolica, cuja maior perda ocorre no produto de fundo da coluna de destilação, ou seja, na vinhaça. Devido à complexidade dessa operação, a literatura apresenta estudos envolvendo ciência de dados. Osório et al. (2008) desenvolveram um sensor virtual para determinar a concentração de álcool destilado a partir de medidas de temperatura utilizando rede neural.

Apesar da maioria dos trabalhos de sensores virtuais na literatura serem com redes neurais, a seleção do método também depende das características do conjunto de dados. Após avaliações, o sensor virtual deste trabalho é baseado na técnica supervisionada denominada k-vizinhos mais próximos (k-NN; k-*Nearest* *Neighbors*), a partir de uma abordagem de problema de classificação. Entre as suas vantagens, destaca-se a implementação simples e a natureza não-paramétrica (BISHOP, 2006).

Um exemplo de sua utilização está no trabalho de Yang et al. (2017), que usaram essa técnica nas variáveis independentes e análise por kernel componentes principais (KPCA; *kernel principal component analysis*) nas variáveis relacionais, compondo um sensor virtual multifuncional para detecção de falhas.

O objetivo do presente trabalho consiste em construir um sensor virtual para aferir o teor de álcool presente na vinhaça em um processo de destilação alcóolica. O estudo de caso utiliza os dados de uma safra de uma indústria do setor sucroenergético no Brasil. A seguir, apresentam-se a metodologia, os resultados e discussões, e a conclusão.

**Metodologia**

O banco de dados inicialmente obtido na indústria é denominado de conjunto de dados crus. Esse conjunto é geralmente coletado independentemente do estado do processo, o que acarreta em valores nulos (*missing values*) e dados anômalos (*outliers*). Com isso, a primeira etapa em análises de dados diz respeito à preparação dos dados. Após a remoção dos dados nulos, aplicou-se o filtro de Hampel para a remoção dos dados discrepantes.

1. Filtro de Hampel

O filtro de Hampel (1971) utiliza a mediana (Med) como estimador de posição central, e o desvio absoluto da mediana (MAD) como estimador de dispersão (Equação (1)). Dado uma variável ($x$), o seu valor ($x\_{i}$) é reconhecido como dado anômalo se está além de$ Med\left(x\right)\pm 3MAD$. Na sequência, um vetor de observações é considerado uma amostra discrepante se pelo menos um de seus valores foi classificado como dado anômalo. Essa alternativa é mais robusta em relação àquela convencional dada por $\overbar{x}\pm 3s$, em que $\overbar{x}$ e s são a média e o desvio-padrão amostrais, respectivamente (Lin et al., 2007).

$MAD=1,4826∙Med\left\{\left|x\_{i}-Med\left(x\right)\right|\right\}$ (1)

1. Conjuntos de dados de treinamento, validação e teste

Um modelo sobreajustado é aquele que explica até o ruído contido no conjunto de dados, o que reduz a sua capacidade de generalização. Já um modelo subajustado não consegue explicar, minimamente, o comportamento da variável de interesse, sendo inadequado para predição (HASTIE et al., 2009).

De modo a mitigar esses problemas, divide-se o conjunto de dados, de forma aleatória, em subconjuntos de identificação (aproximadamente 75% dos registros) e de teste. O primeiro é usado para estimar os parâmetros dos modelos candidatos e para selecionar o modelo final, e o segundo, para verificar a sua capacidade de generalização. Ainda, deve-se garantir que o maior e o menor valor de cada variável estejam no subconjunto de identificação, a fim de se evitar extrapolações (JAMES et al., 2017).

No subconjunto de identificação, aplica-se o procedimento de validação cruzada, que consiste na divisão aleatória dos dados em *m* partições de mesmo tamanho. Utilizou-se *m* = 10 neste trabalho. Em seguida, utilizam-se (*m*-1) subgrupos, denominado conjunto de treinamento, para obter o modelo, e o *m*-ésimo grupo, denominado conjunto de validação, para avaliar a capacidade de predição do modelo (JAMES et al., 2017).

Esse procedimento é repetido *m* vezes, ou seja, para cada partição, usada como grupo de validação. Desse modo, têm-se *m* estimativas da capacidade de predição do modelo, cujo valor médio é computado. Esse procedimento é realizado para cada valor de k, o número de vizinhos mais próximos. para então selecionar o melhor modelo com menor erro médio de predição.

1. k-Vizinhos mais Próximos

O algoritmo k-vizinhos mais próximos (k-NN; k-*Nearest* *Neighbors*) utiliza aprendizado supervisionado. Esse algoritmo classifica um novo ponto a partir do cálculo de suas distâncias a um conjunto de pontos previamente rotulados. Ou seja, a classe de um novo ponto será aquela majoritária entre os k-vizinhos mais próximos ao ponto.

Os parâmetros em k-NN são, o valor de k, que é o número de vizinhos mais próximos; o peso d, ou seja, o grau de influência de cada ponto vizinho sobre o novo ponto, quanto mais distante, menor o peso; o algoritmo de busca, responsável por encontrar os k-vizinhos mais próximos; e a métrica de distância. Neste trabalho, variou-se *k* de 1 até 50; o peso, entre uniforme e baseado na distância; o algoritmo de busca, entre força bruta, *kd-tree* e *ball\_tree* (OMOHUNDRO, 1989); e a métrica de distância, entre Manhattan, Euclidiana e Chebyshev (GOLDBERGER et al., 2005).

Para a seleção do melhor modelo candidato utilizou-se a função custo de entropia cruzada. A função custo é usada não apenas como uma medida de erro, mas também para medir o esforço dispendido para acertar e errar a classificação de cada amostra. Em outras palavras, essa métrica é uma medida do número médio de *bits* necessários para a identificação de um evento de interesse, ou seja, acertar a predição (BISHOP, 2006).

A avaliação final do modelo selecionado utilizou a métrica de acurácia e a matriz confusão. A matriz confusão é uma tabela com as taxas de acertos e erros de cada classe do problema, e a acurácia refere-se ao percentual de classificações corretas (JAMES et al., 2017).

**Resultados e Discussões**

A Figura 1 apresenta o fluxograma do processo usado como estudo de caso. Em vermelho, estão as variáveis medidas em campo por sensores físicos, e em verde, as variáveis medidas em laboratório, a partir de coleta de amostras em campo.

As variáveis selecionadas para a construção do sensor são, (1) a vazão de entrada de vinho na coluna A, (2) a pressão de alimentação do vapor vegetal, (3) a pressão na base da coluna A, (4,5) as temperaturas do vinho após o condensador E e o trocador K, e (6,7) as temperaturas da coluna A no topo e na base.



Figura 1. Esquema simplificado do processo de destilação do bioetanol.

Após a análise do comportamento da variável resposta, selecionou-se a faixa usual de operação entre 0,0 e 0,2 (valores normalizados). Segundo a indústria do estudo de caso, o teor máximo aceitável de álcool na vinhaça é de 0,0121. Assim, a variável resposta foi categorizada em duas classes, conforme a Tabela 1.

Tabela 1. Classes de valores para a variável resposta, o teor de álcool residual na vinhaça.

|  |  |
| --- | --- |
| Classe | Faixa de valores |
| +1 | Y ≤ 0,0121 |
| -1 | Y ≥ 0,0121 |

Após a etapa de tratamento de dados, ou seja, remoção de dados nulos e de dados discrepantes, através do filtro de Hampel, dividiu-se o conjunto de dados em grupo de identificação (75% das observações) e grupo de teste.

Inicialmente, observou-se que os melhores resultados são obtidos com o peso baseado em distância, o algoritmo bruto, e o cálculo da distância pela equação de Chebyshev. A Figura 2 apresenta os valores médios da função custo (à direita) e da acurácia (à esquerda), dado o procedimento de validação cruzada, em função dos valores de *k,* de 1 a 50, para essa combinação de parâmetros. Pode-se observar que a acurácia aumenta com o aumento de *k*, até se estabilizar. Portanto, selecionou-se *k* = 27.



Figura 2. Modelos *k*-NN dado o peso baseado em distância, o algoritmo bruto, e a distância de Chebyshev.

A Tabela 2 contém os valores médios com o desvio da validação cruzada, da função custo e da acurácia para o conjunto de dados de identificação, dado esse modelo final.

A Figura 3 apresenta a matriz confusão resultante da aplicação desse modelo final sobre o conjunto de dados de teste, o que é importante para se verificar a capacidade de predição do modelo.

Tabela 2. Métricas do modelo selecionado, dado o conjunto de dados de identificação.

|  |  |
| --- | --- |
| **Métrica** | **Valor ± desvio** |
| Função Entropia Cruzada | 576,81 ± 43,09 |
| Acurácia | 85,40% ± 1,10 |

Pode-se observar que o sensor virtual obteve uma acurácia de 89,03%, ou seja, aproximadamente 89% das predições foram corretas para dados até então desconhecidos pelo modelo, o que caracteriza um grau significativo de generalização. Pode-se observar também a maior capacidade do modelo em reconhecer a classe +1 (271/(2+274) ≈ 98,9%), o que é esperado pelo seu maior número de observações (classe majoritária) em relação à classe −1 (78/(78+41) ≈ 65,5%). Esse fato constitui-se em um ponto de melhoria.

|  |  |  |
| --- | --- | --- |
| Classe Real |  | Classe Predita |
|  | −1 | +1 |
| −1 | 78 | 41 |
| +1 | 2 | 271 |

Figura 3. Matriz confusão, dado o conjunto de dados de teste.

**Conclusão**

O uso de sensores virtuais em indústrias de processos é importante por permitir o monitoramento contínuo de variáveis-chave, seja por questões de economia, segurança operacional, produção limpa, ou controle de qualidade, como neste trabalho.

Neste trabalho, construiu-se um sensor virtual, baseado no algoritmo k-NN, com o objetivo de prever, de modo contínuo, o teor de álcool residual na vinhaça., Essa informação antecipada é importante em caso de perda de produto na coluna, de modo a agilizar o processo de tomada de decisão do time de operação em direção a normalização do processo. O sensor virtual apresentou uma acurácia significativa de 89%, cuja classificação da classe minoritária, relacionada ao teor de álcool acima daquele desejado, é passível de melhoria.

Por fim, indústrias de processos em geral, e de modo mais específico, o setor sucroenergético, podem se beneficiar de sensores virtuais construídos diretamente a partir de grandes quantidades de dados de operação atualmente à disposição, de modo a aumentar a eficiência das operações.

**Agradecimentos**

Os autores agradecem a indústria sucroenergética pela cessão de um conjunto de dados, e a agência de fomento à pesquisa, CAPES, pelo suporte financeiro.

**Referências Bibliográficas**

ALMEIDA, G. M.; PARK, S. W. Big Data Analytics em Engenharia Química. Revista Brasileira de Engenharia Química, 2º quadrimestre, p. 15– 20, 2017.

BISHOP, C. Pattern Recognition and Machine Learning. 1 ed. New York: Springer-Verlag, 2006

GILBERT, K.; SÀNCHEZ-MARRÈ, M.; RODRIGUEZ-RODA, I. GESCONDA: An intelligent data analysis system for knowledge discovery and management in environmental databases. Environmental Modelling & Software, v. 21, p. 115–120, 2006.

GOLDBERGER, J.; ROWEIS, S.; HINTON, G. Neighbourhood Components Analysis. Advances in Neural Information Processing Systems, v. 17, p. 513-520, 2005.

HAMPEL, F. R. A general qualitative definition of robustness. In: Annals of Mathematics Statistics, 42, p. 1887–1896, 1971

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction. 2 ed. Springer-Verlag, 2009.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning with Applications in R. 8. ed. New York: Springer, 2017

LIMA, R. N.; DE ALMEIDA, G. M.; BRAGA, A. P.; CARDOSO, M. Trend modelling with artificial neural networks. Case study: Operating zones identification for higher SO3 incorporation in cement clinker. Engineering Applications of Artificial Intelligence, v. 54, p. 17-25, 2016.

LIN, B.; RECKE, B.; KNUDSEN, J. K. H.; JORGENSEN, S. B. A systematic approach for soft sensor development. Computers and Chemical Engineering, v. 31, p. 419-425, 2007.

OMOHUNDRO, S. M. Five ball-tree construction algorithms. International Computer Science Institute, ICSI Technical Report TR-89-063, 1989

OSORIO, D.; PÉREZ-CORREA, J. R.; AGOSIN, E.; CABRERA, M. Soft sensor for on-line estimation of ethanol concentrations in wine stills. Journal of Food Engineering, v. 87, p. 571–577, 2008.

YANG, J.; LIN, L.; SUN, Z.; CHEN, Y.; JIANG, S. Data Validation of Multifunctional Sensors Using Independent and Related Variables. Sensors and Actuators A: Physical. v. 263, p. 76-90, 2017.