Classificadores de desempenho nos processos de mistura e aquecimento de uma planta industrial didática

Performance classifiers in the mixing and heating processes of a didactic industrial plant

Guilherme da Cunha (orientado)[1], Gláucia Maria Bressan (orientadora)[2], Wagner Endo[3]

RESUMO

O objetivo deste trabalho é a aplicação e análise de métodos de Machine Learning para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática localizada na UTFPR do câmpus Cornélio Procópio. Os métodos aplicáveis para esta classificação são k-nearest neighbors (KNN), Árvores de Decisão, Florestas Randômicas e Naive- Bayes. Para esta tarefa, as variáveis de entrada consideradas são a porcentagem de abertura da válvula, a vazão e o tempo da abertura; e a variável de saída é a temperatura, discretizada em 5 classes. O desempenho dos algoritmos é analisado considerando-se a acurácia e medidas estatísticas relevantes e as implementações dos métodos são feitas utilizando o software R. Os resultados apresentam um bom desempenho dos algoritmos na tarefa de classificação da temperatura do processo de mistura de líquidos, com acurácias acima de 90%.

Palavras-chave: machine learning, aquecimento, temperatura, processo de mistura.

ABSTRACT

The goal of this paper is the application and analysis of Machine Learning methods to classify the temperature resulting from the liquids mixing process in a didactic industrial plant located in the UTFPR of the Cornélio Procópio city. The applicable methods for this classification are k-nearest neighbors (KNN), Decision Trees, Random Forest and Naive- Bayes. For this task, the input variables considered are the percentage of valve opening, the flow and the opening time; and the output variable is temperature, discretized into 5 classes. The performance of the algorithms is analyzed considering the accuracy and relevant statistical measures and the implementations of the methods are made using the R software. The results show a good performance of the algorithms in the task of classifying the temperature of the liquid mixing process, with accuracy above 90%.

Keywords: machine learning, heating, temperature, mixing process

1 INTRODUÇÃO

Os métodos de Machine Learning, ou “aprendizado de máquina”, são métodos de análise de dados provenientes do estudo da inteligência artificial e têm como principal objetivo automatizar a construção de

[1] Engenharia Mecânica, Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio; guilherme_cunha20@hotmail.com
[2] Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio; glauciabressan@utfpr.edu.br
[3] Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio; wendo@utfpr.edu.br

modelos analíticos. A ideia é que o sistema possa analisar os dados conhecidos, identificando padrões, e com isso, possa tomar decisões sem a intervenção humana (AGGARWAL, 2015).

Neste trabalho, são aplicados diferentes métodos de Machine Learning para análise do desempenho dos algoritmos na tarefa de classificar a temperatura no processo de mistura de líquidos de uma planta industrial didática localizada na Universidade Tecnológica Federal do Paraná, campus Cornélio Procópio.

Neste contexto, o objetivo deste trabalho é a aplicação e a análise de métodos de Machine Learning para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática localizada na UTFPR do câmpus Cornélio Procópio. Os métodos considerados para esta classificação são KNN, Árvores de Decisão, Florestas Randômicas e Naive- Bayes. Para isso, as variáveis de entrada consideradas são a porcentagem de abertura da válvula, a vazão e o tempo da abertura; e a variável de saída é a temperatura, a qual é discretizada em 5 classes (muito baixa, baixa, média, alta e muito alta). O desempenho dos algoritmos é analisado considerando-se a acurácia e medidas estatísticas relevantes, e as implementações dos algoritmos são feitas utilizando o software R (https://cran.r-project.org/bin/ windows/ base/).

2 MÉTODO

O trabalho foi desenvolvido utilizando a planta didática SMAR (SMAR, 2015), que reproduz processos industriais para fins didáticos, podendo simular experimentos muito próximo de situações reais.

A planta didática simula dois processos, sendo eles, aquecimento e mistura. O aquecimento ocorre através de uma resistência de submersão, instalada no tanque de aquecimento. Já a mistura é feita a partir do líquido aquecido e do frio, proveniente do outro reservatório. Este é acionado a partir de um controlador, e responde as variações de temperatura do tanque quente, mantendo a mistura na temperatura desejada (setada previamente).

O sistema identificará a demanda de água fria, e a utilizará para manter a temperatura desejada, a misturando com a água quente. O processo procura o equilíbrio térmico desejado de forma empírica.

Neste trabalho, optou-se por utilizar o software R para implementação dos classificadores, visto que este já possui os pacotes necessários. Primeiramente, é necessário realizar o pré-processamento dos dados, por meio do tratamento de dados faltantes e outliers, e selecionando as variáveis de entrada do problema. É importante notar que o pré-processamento é sempre o primeiro passo a ser realizado, independente do modelo escolhido para realizar a tarefa de classificação.

Após realizado o pré-processamento dos dados, pode-se iniciar a tarefa de classificação por meio dos algoritmos de Machine Learning. Neste trabalho, os modelos aplicáveis para a classificação da temperatura no processo de mistura na planta industrial são: KNN, Árvores de Decisão, Florestas Randômicas e Naive- Bayes (AGGARWAL, 2015).

3 RESULTADOS E DISCUSSÃO

O conjunto de dados contém 1202 linhas, que consistem nos valores numéricos das 3 variáveis de entrada (percentual de abertura da válvula, tempo e vazão) e também o valor numérico de saída (Temperatura). Todos os dados estão normalizados, ou seja, variam de 0 a 1.

Para realizar a classificação, primeiramente, é necessário discretizar a variável de saída, Temperatura, de modo hierárquico. De acordo com Bressan; Mosaner; Endo, (2020), foram escolhidas cinco classes de temperatura: 1- muito baixa, 2- baixa, 3- média, 4- alta e 5- muito alta.

O conjunto de dados é dividido nos subconjuntos de treinamento (treinamento do modelo, que consiste em "aprender" a classificar corretamente a variável de saída) e de teste (utilizado para verificação de sua acurácia, ou índice de acerto da classificação), de acordo com o método de validação cruzada com n = 10 (AGGARWAL,2015).

O método KNN é implementado no software R com o auxílio das bibliotecas "caTools", "class" e "caret", de forma que seja possível realizar a classificação e obter a matriz de confusão, a fim de verificar o desempenho da classificação. Além disso, os resultados estatísticos auxiliam para a análise do desempenho do método.

A matriz de confusão para o método KNN utilizando um número de vizinhos k = 9 é dada a seguir.

classes	**1**	**2**	**3**	**4**	**5**
1	26	0	0	0	0
2	0	20	0	0	0
3	0	0	44	2	0
4	0	0	0	82	2
5	0	0	0	1	124

Vale ressaltar que o número de vizinhos a ser utilizado foi obtido de forma iterativa, simplesmente testando valores comuns de utilização, a começar por k=5. O método se mostrou extremamente eficaz para o problema analisado, com apenas 5 classificações erradas e nenhuma classificação discrepante. Como é possível notar resquícios do aumento de temperatura quando o tempo de vazão aumenta, era esperado obter mais erros em classes de temperatura mais alta, mesmo assim, a maioria dos valores encontra-se na diagonal principal da matriz, a qual representa os acertos do método.

Para a verificação da eficácia do KNN como classificador para o problema apresentado, utilizou-se medidas estatísticas, das quais destaca-se as mais importantes: eficácia de 0,9834, intervalo de confiança de 0,9617 a 0,9946 e Kappa de 0,9767. Nota-se a alta acurácia, indicando um índice de acerto acima de 98%, além disso, o valor de Kappa mostra uma alta concordância entre os dados. As definições para as estatísticas apresentadas podem ser encontradas em Aggarwal (2015). Vale ressaltar que foram utilizadas duas rodadas de validação cruzada para diferentes valores de K, além disso, mesmo que os melhores resultados sejam provenientes de K=9, para outros valores de K, a acurácia permaneceu acima de 98%, indicando a robustez do método.

Para a implementação das Árvores de Decisão, são utilizadas as seguintes bibliotecas: "caTools", "rpart", "rpart.plot" e "caret". A matriz de confusão é apresentada a seguir.

classes	1	2	3	4	5
1	26	0	0	0	0
2	0	20	0	0	0
3	0	0	46	2	0
4	0	0	1	82	1
5	0	0	0	1	124

Ao observar a matriz, nota-se o bom desempenho do classificador, visto que quase todos os valores se encontram na diagonal principal, a qual representa os acertos. O algoritmo se mostra eficiente mesmo com a dificuldade de se classificar as altas temperaturas provenientes de aquecimentos anteriores. Destaca-se também as principais medidas estatísticas obtidas: acurácia de 99%, intervalo de confiança de 0,9712 a 0,9979 e Kappa de 0,9861.

Para a implementação das Florestas Randômicas, foram utilizadas as seguintes bibliotecas: “caTools”, “randomForest”, “caret”. Como citado anteriormente, esse algoritmo permite uma confiabilidade maior do que as árvores de decisão, pois elimina ou ao menos diminui significativamente a chance de as estatísticas serem influenciadas por seleções boas ou ruins de conjunto de treinamento e teste. A matriz de confusão obtida para o método das florestas é mostrada a seguir.

classes	1	2	3	4	5
1	26	0	0	0	0
2	0	20	0	0	0
3	0	1	45	0	0
4	0	0	1	77	6
5	0	0	0	1	124

O desempenho do classificador pode ser observado na matriz de confusão, visto que a maioria dos valores se encontram na diagonal principal, que representa os acertos da classificação. Nesse caso, obteve-se um valor de 97,01% de eficácia, um intervalo de confiança de 0,944 a 0,986 e um Kappa de 0,9581. O alto desempenho do classificador pode ser notado pelos valores de acurácia e intervalo de confiança. A acurácia é um pouco menor do que o método das Árvores de Decisão, justamente porque o modelo utiliza uma combinação de 100 árvores para obter tais resultados, o que garante maior confiabilidade do método. O valor do índice Kappa mostra uma boa concordância dos dados. Por fim, foram realizadas duas rodadas de validação cruzada para o método. O retorno foi de uma acurácia de 0,9929200 e Kappa de 0,9900797, confirmando o bom desempenho do classificador.

Para a implementação do classificador Naive Bayes, são utilizadas as seguintes bibliotecas: “caTools”, “e1071” e “caret”. Dessa forma, verifica-se os resultados da classificação com o auxílio da matriz de confusão e das estatísticas descritivas. A matriz de confusão é apresentada a seguir.

classes	1	2	3	4	5
1	26	0	0	0	0
2	0	20	0	0	0
3	0	5	39	2	0
4	0	0	0	74	10
5	0	0	0	11	114

A maioria dos valores se encontra na diagonal principal, mostrando um bom desempenho do classificador, além disso, não há classificações muito discrepantes. Novamente, a maior dificuldade se encontra em classificar as temperaturas mais altas, visto que há resquícios dos aquecimentos anteriores, dificultando a caracterização do comportamento. Para esse método, o valor de acurácia obtido foi de 90,7%, um intervalo de confiança de 0,8684 a 0,9373 e um Kappa de 0,87. A acurácia de 90,7% mostra o bom desempenho do método *Naive Bayes*, apesar de outros classificadores citados se mostrarem ainda melhores.

Para gerar uma maior confiabilidade e evitar medidas ao acaso, utilizou-se duas rodadas de validação cruzada, obtendo-se uma acurácia de 0,9433408 e valor do índice Kappa de 0,9201878. Isso mostra que as condições do primeiro teste (por exemplo a seleção de dados de treinamento) não era tão favorável, e que em geral, o método tem um desempenho melhor do que o mostrado anteriormente.

Comparando os resultados das estatísticas descritivas, juntamente com as matrizes de confusão de cada método e considerando as validações cruzadas, nota-se primeiramente que o classificador Naive Bayes apresentou pior desempenho na classificação em comparação aos outros métodos para o caso apresentado. Isso pode ser justificado pela simplicidade do método e baixíssimo custo computacional envolvido no processamento do mesmo. É importante ressaltar que apesar de comparativamente os resultados para o método Naive Bayes serem os piores, em geral, a acurácia de aproximadamente 94% é alta, mostrando que a ferramenta possui um bom custo-benefício.

Em relação à acurácia, os maiores valores (99% ou acima) foram obtidos pelo método das Árvores de Decisão e Florestas Randômicas, entretanto, a acurácia de 98,58% obtida pelo KNN também deve ser notada, visto que o custo computacional envolvido é menor. As Florestas Randômicas apresentaram o maior valor de Kappa, mostrando uma melhor concordância entre os dados, porém, ambos os métodos KNN e Árvores de Decisão também apresentaram valores elevados (acima de 98%). Vale lembrar também que as Florestas Randômicas apresentam resultados mais confiáveis, entretanto, o custo computacional envolvido na criação das florestas é extremamente maior.

Em especial, o método KNN se destaca pela sua constância, mantendo valores extremamente elevados de acurácia, intervalo de confiança, Kappa, sensitividade e especificidade à um custo computacional relativamente baixo quando comparado com as Árvores de Decisão, que apresenta os melhores valores brutos das estatísticas descritivas, porém, a um custo computacional maior e menor confiabilidade e robustez, quando comparado às Florestas Randômicas.

É importante ressaltar que as Árvores de Decisão têm a vantagem na questão da interpretação dos resultados, apresentando um resultado visual ao usuário que pode ser visto como decisões lógicas, facilmente verificadas. Vale ressaltar que o alto custo das Florestas Randômicas e Árvores de Decisão está associado ao processo de criação das árvores, porém, uma vez criadas, o custo é baixo para a utilização dos métodos.

4 CONCLUSÃO

Este trabalho propõe aplicação e análise de métodos de *Machine Learning* para a classificação da temperatura resultante do processo de mistura de líquidos de uma planta industrial didática, localizada na UTFPR do câmpus Cornélio Procópio.Todos os métodos considerados apresentaram bons resultados em relação à acurácia e às medidas estatísticas. O método Naive Bayes tem bons resultados a um menor custo computacional, enquanto as Árvores de Decisão mostram as melhores estatísticas descritivas e maior quantidade de acertos a um alto custo computacional. Ao mesmo tempo, as Florestas Randômicas apresentam a maior confiabilidade e robustez, porém, é o método com maior custo computacional, em relação aos demais. Portanto, os classificadores apresentados se mostram como sendo ferramentas eficazes para previsão de dados e com ótimo desempenho associado à tarefa de classificação.

Como perspectivas de continuidade do trabalho, propõe-se uma nova coleta de dados na planta industrial didática, para elaboração de métodos de Machine Learning para a classificação de falhas dos processos executados pela planta em estudo. Tais falhas podem ser classificadas em determinados grupos e o desempenho dos algoritmos podem ser analisados de forma similar.

REFERÊNCIAS

AGGARWAL, C. C. (Ed.). **Data Classification**: algorithms and applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Book 35, 2015.

BRESSAN, G. M.; SILVA, G. M.; ENDO, W**.** Estratégia para Compensação de Erros de Ação de Controle em uma Válvula Industrial Utilizando Inferência Fuzzy. **Revista de Engenharia e Tecnologia**. v.12, p.223 - 234, 2020.

SMAR. **Manual de instruções, operação e manutenção**: plantas didáticas. PD3-P, versão 3, 2015.