Anais 2025
27ª Semana de Pesquisa da Universidade Tiradentes
“Ciência, Água e Vida: Inovação para transformar o clima do amanhã”
03 a 07 de novembro de 2025
ISSN: 1807-2518

UTILIZAÇÃO DE ALGORÍTMOS DE MACHINE LEARNING PARA O DESENVOLVIMENTO DE MODELOS DE CLASSIFICAÇÃO E QUANTIFICAÇÃO DE BANCO DE DADOS COM ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO

Allan Bruno Santana Prado (bolsista Petrobras/ITP);
Raphael José Laurindo Santana (Voluntário);
Ayslan Santos Pereira da Costa (Orientador);
Cláudio Dariva (Orientador);
Gustavo Rodrigues Borges (Orientador);

allan.bruno@souunit.com.br;

Universidade Tiradentes/Ciências da Computação/Aracaju/SE.
Instituto de Tecnologia e Pesquisa - ITP

1.00.00.00-3 - Ciências Exatas e da Terra; 1.03.00.00-7 - Ciência da Computação

RESUMO

Algoritmos de aprendizado de máquina (ML) têm se destacado em diversas aplicações industriais, especialmente no monitoramento de propriedades físico-químicas. Entre essas aplicações, destaca-se a predição de teores de monossacarídeos, como glicose, xilose, galactose, manose, arabinose e celobiose, obtidos em diferentes processos que utilizam biomassa lignocelulósica como matéria-prima para produção de biocombustíveis. Atualmente, a técnica mais comum para avaliar a eficiência desses processos são os métodos cromatográficos, que requerem alto custo, preparo de amostra e elevado tempo de análise. Nesse contexto, a espectroscopia de infravermelho próximo (NIR) surge como alternativa promissora para o monitoramento de processos, por possibilitar análise *in-line* com sondas resistentes a condições severas comumente encontradas na indústria. Contudo, os espectros gerados pelo NIR são complexos, exigindo modelos matemáticos e ferramentas computacionais capazes de converter as respostas do equipamento em informações relevantes sobre o analito.O presente estudo tem como objetivo desenvolver e validar um modelo de aprendizado de máquina para quantificar e classificar açúcares em soluções aquosas. Para isso, utilizaram-se dados obtidos em um espectrômetro FT-NIR Bruker, modelo MPA®, com técnica de transmitância, em soluções aquosas de padrões de açúcares em concentrações de 100 a 5000 ppm, totalizando 295 espectros, sendo cinco para cada concentração. O conjunto de dados foi organizado em classes correspondentes ao tipo de açúcar. Após a obtenção e organização, os espectros foram pré-processados utilizando a técnica SNV, destacando o efeito da concentração e do tipo de açúcar no sinal espectral.Para a classificação dos açúcares, foi empregado um modelo CNN1D, uma rede convolucional projetada para processar dados sequenciais, como espectros NIR. A CNN1D foi definida com uma camada de convolução com kernel de tamanho 3 e função de ativação ReLU, garantindo a não linearidade do aprendizado. Uma camada de *Max Pooling* reduziu a dimensionalidade dos dados, permitindo à rede aprender características predominantes. Em seguida, uma camada densa previu o tipo de açúcar utilizando a função Softmax.As amostras contendo glicose foram submetidas a ajuste de modelo de regressão. Nesta etapa, aplicou-se seleção de região espectral, o método Detrend para eliminar tendências lineares e Boxplot para identificar e remover *outliers*. O algoritmo Partial Least Squares Regression (PLSR) foi selecionado, configurado com cinco componentes principais. Em ambas as etapas (classificação e regressão), os dados foram divididos em 70% para treinamento e 30% para teste. O desempenho foi avaliado por meio da função

de perda, acurácia, matriz de confusão e validação cruzada.Os modelos ajustados apresentaram resultados promissores: a classificação atingiu 99% de acurácia para glicose, xilose, arabinose e galactose; e o modelo de regressão obteve média de Coeficiente de determinação (R^2) de 96% e Raiz do erro quadrático médio (RMSE) de 297,8 para glicose. Esses resultados indicam que os modelos são promissores em soluções aquosas contendo um tipo de monossacarídeo. Trabalhos futuros buscarão adaptar os modelos para reconhecer amostras mais complexas e prever concentrações de múltiplos açúcares.

PALAVRAS-CHAVE: Espectroscopia de infravermelho próximo, aprendizado de máquinas, processamento de dados

ABSTRACT

Machine learning (ML) algorithms have gained prominence in various industrial applications, especially in monitoring physicochemical properties. Among these applications, the prediction of monosaccharide contents such as glucose, xylose, galactose, mannose, arabinose, and cellobiose stands out in processes that use lignocellulosic biomass as raw material for biofuel production. Currently, the most common technique for evaluating the efficiency of these processes is chromatography, which involves high costs, extensive sample preparation, and long analysis times. In this context, near-infrared (NIR) spectroscopy emerges as a promising alternative for process monitoring, as it allows in-line analysis using probes resistant to the harsh conditions commonly found in industrial environments. However, NIR spectra are complex and require mathematical models and computational tools capable of converting instrument responses into meaningful information about the analyte.This study aims to develop and validate a machine learning model for quantifying and classifying sugars in aqueous solutions. Data were obtained using a Bruker FT-NIR spectrometer, model MPA®, operating in transmittance mode, from aqueous sugar standard solutions with concentrations ranging from 100 to 5000 ppm, totaling 295 spectra, with five spectra recorded for each concentration. The dataset was organized into classes corresponding to each sugar type. After acquisition and organization, spectra were preprocessed using the Standard Normal Variate (SNV) technique, highlighting the effects of concentration and sugar type on the spectral signal.For sugar classification, a one-dimensional convolutional neural network (CNN1D) was employed, designed to process sequential data such as NIR spectra. The CNN1D was defined with a convolutional layer using a kernel size of 3 and a ReLU activation function to ensure nonlinear learning. A Max Pooling layer reduced data dimensionality, allowing the network to learn predominant features. Subsequently, a dense layer predicted the sugar type using the Softmax activation function.Samples containing glucose were subjected to regression model fitting. In this step, spectral region selection was applied, along with the Detrend method to remove linear trends and the Boxplot method to identify and remove outliers. The Partial Least Squares Regression (PLSR) algorithm was selected and configured with five principal components. In both classification and regression stages, data were split into 70% for training and 30% for testing. Model performance was evaluated using loss function, accuracy, confusion matrix, and cross-validation.The adjusted models showed promising results: the classification achieved 99% accuracy for glucose, xylose, arabinose, and galactose, while the regression model achieved an average coefficient of determination (R^2) of 96% and a root mean square error (RMSE) of 297.8 for glucose. These results indicate that the proposed models are effective for aqueous solutions containing a single type of monosaccharide. Future work will focus on adapting the models to recognize more complex samples and to predict concentrations of multiple sugars.

KEYWORDS: Near-infrared spectroscopy, machine learning, data processing