Distance-based applicability domain for screening Cu-based tantalate materials for visible-light-responsive photocatalyst candidates

Mayumi Puspita
Nara Intitute of Science and Technology

Materials informatics integrates machine learning and materials science to accelerate the classification and discovery of functional materials. This study introduces a supervised learning framework to distinguish between zero and non-zero band gaps using a distance-based Applicability Domain (AD) strategy. The AD concept assesses prediction reliability by evaluating the closeness between unseen data and the training data in a reduced-dimensional space. Principal Component Analysis (PCA) reduces feature complexity, while k-nearest neighbors (k-NN) determine local class composition and spatial proximity. This combination allows the model to focus on dense regions with uniform neighbor labels and reduce ambiguity in sparse areas. Both the confidence score and the weighted label ratio quantify the reliability of the prediction. Visualization of the weighted ratio reveals domain boundaries and highlights irregular prediction behavior. This strategy helps the model avoid misleading classifications outside the reliable zone. It also facilitates a clearer interpretation of model output based on local data structure. The method enhances screening accuracy across complex chemical systems by combining statistical and structural information. Focusing on Cu-based tantalate photocatalyst materials as candidates for hydrogen production under visible light, the framework offers a robust and scalable approach to domain-aware screening in real-world applications.