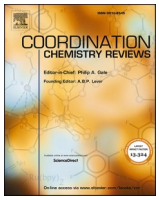
Coordination Chemistry Reviews 536 (2025) 216699

Contents lists available at ScienceDirect

Coordination Chemistry Reviews

journal homepage: www.elsevier.com/locate/ccr



Advancing robust clustering verification for unsupervised photocatalytic data analysis

ARTICLE INFO

Keywords:
Unsupervised learning
Clustering validation
HDBSCAN
OPTICS
Feature identification
Robust statistical methods

ABSTRACT

Recent research by Mahapatra et al. highlights the promise of unsupervised learning techniques in 2D MXenes-based photocatalytic applications; however, the absence of ground truth data poses significant challenges in validating both feature identification and clustering quality. The approach presented here advocates for the integration of advanced clustering methods that overcome the limitations of traditional techniques. In particular, nonlinear and nonparametric algorithms, such as HDBSCAN and OPTICS, are favored for their ability to accommodate irregular data structures without relying on conventional clustering assumptions. Additionally, complementary evaluation metrics—including the Silhouette Score, Davies-Bouldin Index, and Gap Statistic—are introduced to comprehensively assess cluster cohesion, separation, and the optimal number of clusters. This integrated framework is designed to enhance the validation of unsupervised clustering outcomes and improve the overall reliability of analyses in photocatalytic research.

Mahapatra et al. investigated the role of artificial intelligence in 2D MXenes-based photocatalytic applications [1]. They highlighted the use of unsupervised learning techniques, which facilitate classification and grouping within a dataset based on similarities, thereby identifying the most relevant information. Key methods in unsupervised learning include K-means clustering, autoencoders, and generative adversarial networks (GANs) [1].

The core challenge in unsupervised machine learning is the lack of ground truth labels, which makes validation significantly more difficult compared to supervised learning. This challenge is evident in studies such as Mahapatra et al.'s, where the approach to cluster validation is not clearly outlined. In contrast to supervised methods that can be directly evaluated against known outcomes, unsupervised techniques must depend on internal validation metrics—a predicament that nonlinear and nonparametric robust statistical methods attempt to mitigate [2,3].

Conventional clustering methods, like K-means, rely on several key assumptions: clusters are spherical, variances are uniform, distances are measured in Euclidean space, and boundaries are linear. These assumptions frequently fail when applied to complex, nonlinear, and nonparametric data in coordination chemistry, where clusters often display irregular shapes and varying densities. As a result, the use of traditional methods can lead to distorted and potentially misleading conclusions [4–8].

To overcome these issues, we advocate for the use of advanced clustering algorithms: HDBSCAN [9] and OPTICS [10]. HDBSCAN builds upon DBSCAN through hierarchical clustering, which allows it to handle clusters of varying densities and shapes while autonomously identifying the optimal number of clusters. OPTICS, on the other hand, creates a density-based ordering of points, facilitating the detection of clusters across different density scales without the need for manual parameter adjustments.

For a rigorous evaluation of clustering performance, we employ three complementary metrics: the Silhouette Score [11], which assesses both cluster cohesion and separation; the Davies-Bouldin Index [12], which measures the similarity between clusters; and the Gap Statistic [13], which objectively determines the optimal number of clusters by comparing the observed within-cluster variation to that expected under a null reference distribution. This integrated evaluation framework offers a systematic approach to validating clustering outcomes, making it particularly valuable for the analysis of complex chemical datasets.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Not applicable.

Funding

This research has no fund.

Availability of data and material

Not applicable.

https://doi.org/10.1016/j.ccr.2025.216699
Received 6 March 2025; Accepted 2 April 2025

Authors' Contribution

Yoshiyasu Takefuji completed this research and wrote this article.

According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] D.M. Mahapatra, A. Kumar, R. Kumar, N.K. Gupta, B. Ethiraj, L. Singh, Artificial intelligence interventions in 2D MXenes-based photocatalytic applications, Coord. Chem. Rev. 529 (2025) 216460, https://doi.org/10.1016/j.ccr.2025.216460.
[2] Y. Takefuji, Enhancing machine learning in gas–solid interaction analysis: addressing feature selection and dimensionality challenges, Coord. Chem. Rev. 534 (2025) 216583, https://doi.org/10.1016/j.ccr.2025.216583.
[3] Y. Takefuji, Reevaluating feature importance in gas–solid interaction predictions: a call for robust statistical methods, Coord. Chem. Rev. 534 (2025) 216584, https://doi.org/10.1016/j.ccr.2025.216584.
[4] Y.P. Raykov, A. Boukouvalas, F. Baig, M.A. Little, What to do when K-means clustering fails: a simple yet principled alternative algorithm, PLoS One 11 (9) (2016) e0162259, https://doi.org/10.1371/journal.pone.0162259.
[5] C. Wongoutong, The impact of neglecting feature scaling in k-means clustering, PLoS One 19 (12) (2024) e0310839, https://doi.org/10.1371/journal.pone.0310839.
[6] Y.T. Chen, D.M. Witten, Selective inference for k-means clustering, J. Mach. Learn. Res. 24 (2023) 152.
[7] H. Shen, S. Bhamidi, Y. Liu, Statistical significance of clustering with multidimensional scaling, J. Comput. Graph. Stat. 33 (1) (2024) 219–230, https://doi.org/10.1080/10618600.2023.2219708.
[8] J.P. Burgard, C. Moreira Costa, M. Schmidt, Robustification of the k-means clustering problem and tailored decomposition methods: when more conservative means more accurate, Ann. Oper. Res. 339 (2024) 1525–1568, https://doi.org/10.1007/s10479-022-04818-w.
[9] T. Lee, Y. Kim, Y. Hyun, et al., Unsupervised anomaly detection process using LLE and HDBSCAN by style-GAN as a feature extractor, Int. J. Precis. Eng. Manuf. 25 (2024) 51–63, https://doi.org/10.1007/s12541-023-00908-2.
[10] A. Hajihosseinlou, A. Maghsoudi, R. Ghezelbash, A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins, Geochemistry 84 (2) (2024) 126094, https://doi.org/10.1016/j.chemer.2024.126094.
[11] M. Shutaywi, N.N. Kachouie, Silhouette analysis for performance evaluation in machine learning with applications to clustering, Entropy 23 (6) (2021) 759, https://doi.org/10.3390/e23060759.
[12] F. Ros, R. Riad, S. Guillaume, PDBI: a partitioning Davies-Bouldin index for clustering evaluation, Neurocomputing 528 (2023) 178–199, https://doi.org/10.1016/j.neucom.2023.01.043.
[13] I.K. Khan, H.B. Daud, N.B. Zainuddin, et al., Determining the optimal number of clusters by enhanced gap statistic in K-mean algorithm, Egypt. Inform. J. 27 (2024) 100504, https://doi.org/10.1016/j.eij.2024.100504.

Yoshiyasu Takefuji
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.