The Eleventh International Conference on
Engineering Computational Technology
23–25 August, 2022 | Montpellier, France

Proceedings of the Eleventh International Conference on
Engineering Computational Technology
Edited by B.H.V. Topping and P. Iványi
Civil-Comp Conferences, Volume 2, Paper 7.1
Civil-Comp Press, Edinburgh, United Kingdom, 2022, doi: 10.4203/ccc.2.7.1

Interpretable Feature Extraction for the Numerical Particle System

Shuai Ren[1], Xinyu Zhang[1], Huizhao Li[1], Genshen Chu[1], Dandan Chen[1], He Bai[1], Changjun Hu[1,2]

[1]University of Science and Technology Beijing, Beijing, China
[2]Engineering Research Center of Intelligent Supercomputing, Ministry of Education, Beijing, China

Abstract

Particle system analysis is important but costly in solving many physical problems since particles are the basic composition of almost everything, and machine learning is increasingly used in optimization for numerical simulations. The most important and difficult job is to make the particle system understandable by the machine learning model, which usually called feature extraction. In this paper, a novel method and an accurate physical interpretation for feature extraction of cascade defects data, an important particle system in reactor pressure vessel analysis, was proposed. Four strategies were designed to extract features from cascade defects data in which the DAPP shows the best performance. This study shows that feature extraction based on physical information has a positive significance for the analysis of particle systems and provide a theoretical support for improving the physical interpretability of machine learning models on particle systems.

Keywords: particle system, feature extraction, machine learning, numerical simulation, physical interpretability, reactor pressure vessel

1 Introduction

Particle system analysis is important in solving many physical problems since particles are the basic composition of almost everything, like structural materials, medicine, fuel rods, etc. [1-2]. Many years past, numerical simulation has become the most important method in study of particle systems due to its efficiency, safety and cost saving [3-4]. The classic numerical particle models include molecular dynamics, neutron physics, etc. However, the calculation of complex nonlinear problems and partial differential equations becomes the main bottleneck of the development of numerical simulation with larger scale and higher precision [5-7]. The conventional practice is to further simplify the model to achieve a rough simulation at the expense of accuracy, which makes a large gap with practical applications. Machine learning makes things easy due to its powerful fitting ability in solving complex nonlinear problems and partial differential equations, as we can see in many fields like natural language recognition and image recognition [8-10]. In particle systems, the most important and difficult job is to make the particle system understandable by the machine learning model, which called feature extraction, and a set of interpretable features will make the model more accurate and instructive for understanding the physical process [11-12].

This paper presents a method for feature extraction from the perspective of physical interpretability with the cascade defects, a representative numerical particle system in reactor pressure vessel analysis. Specifically, given the cascade defects, simulated by the molecular dynamics, our method aims to extract key physical features for mining of the cascade defects clusters. Different physical features was proposed and a physical explanation was given for their effect as to why these features can get a different result for cascade defects recognition.

2 Methods

In this part, the methods employed in this paper are introduced as follows including data preparation, feature extraction method, clustering algorithm and model validation. **Data Preparation:** The cascade defects data was obtained by MISA-MD, the fastest molecular dynamic simulation software for simulating radiation cascade in reactor pressure vessels developed by USTB [13], and the open source address is https://hpcde.github.io/p/open_source/. The atoms are all of Fe with a BCC crystal structure and the lattice constant is 2.85532. The temperature is 600K and the box size is [80, 80, 80], which means that the size of xyz direction is 80 times the lattice constant. The result of the last time step was selected which was composed of xyz coordinate of 1,024,000 atoms. Different energies and directions of neutron were simulated including 3 energies of 10kev, 30kev, 50kev with three directions of 122, 135, 235 respectively, each of which has been simulated 50 times. Wigner–Seitz algorithm [14] was employed to recognize the interstitial atoms and vacancy from the last step simulation data and Union-Find algorithm [15] was employed to divide the interstitial atoms and vacancy into defects clusters, which are generally regarded as spherical particles. Finally, 4483 clusters were obtained, as shown in Figure 1. Our goal is to classify the clusters that are similar among these 4483 cascade defects data through clustering algorithm.

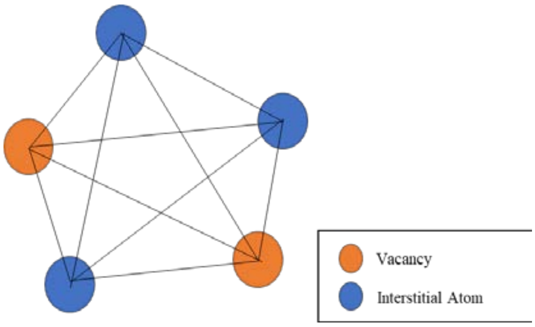

Figure 1 The structure of cascade defects

Feature Extraction Method: Four strategies are designed to extract features from cascade defects data, as shown in the Table 1. DPC represents particle-to-Center of mass distance, where the center of mass here is calculated from all particles in the cluster, and DPP represents particle-to-particle distances. DAPC and DAPP consider one more feature of angle than DPC and DPP.

Features	**Particle-Center of mass**	**Particle-Particle**
Distance	DPC	DPP
Distance + Angle	DAPC	DAPP

Table 1 Four strategies to extract features

Clustering Algorithm: HDBSCAN [16] was employed as the clustering algorithm to the cascade defects data. HDBSCAN is a combination of DBSCAN algorithm [17] and hierarchical clustering algorithm. The algorithm has the same performance as the DBSCAN clustering algorithm and more intuitive parameters. Besides, the HDBSCAN algorithm can deal with clustering problems with different densities more effectively, which is the best choice for clustering cascade defects data.

Model Validation: Silhouette Coefficient [18] was employed to evaluate the quality of the feature strategy, which is a most common evaluation methods used in clustering model.

3 Results

The score of Silhouette Coefficient and noise are listed in the Table 2. As shown in the table, the DAPP strategy shows the best performance with a highest Silhouette Coefficient and a lowest noise value. Specifically, DPP strategy shows a much higher score than DPC strategy when take the particle-to-particle distance as the only feature with an almost identical but very high noise value. That is because features that only contain distances can only roughly describe the density of clusters, but not feasible to describe the morphology of clusters very well, so these two strategies show very poor performance in recognizing cascade defects. The score of Silhouette Coefficient of DPAC has been greatly improved and the noise value has dropped a lot when not only distance but angle are considered. The noise value of DAPP has dropped further when take particle-to-particle distance and angle rather than particle-to-center of mass distance and angle as the feature compared with DAPC.

Feature Strategy	Silhouette Coefficient	Noise (unrecognized clusters)
DPC	0.44	868
DPP	0.71	860
DAPC	0.71	219
DAPP	0.71	111

Table 2 The score of Silhouette Coefficient and noise with different feature strategies

Schematic diagram of the DAPC and DAPP with an example cascade defects was shown in Figure 2, which are two similar clusters, (a) and (c) show the same pentagon-shaped clusters and (b) and (d) show another cluster which add a tail consist of particle 6 and particle 7 to the pentagon-shaped cluster. From the perspective of the two cluster morphology, they have strong similarities which are usually divided into one category. Under the DAPC strategy, the distance and angle values of these two similar clusters are completely different, while only the features of the tail part (particle 6 and particle 7) change with exactly the same values in terms of the pentagon part when the DAPP strategy is used. Consequently, such clusters can be classified into the same category. From the point of view of the physical information, the analysis above provides an accurate explanation for the pros and cons of these feature strategies to recognize the cascade defect system.

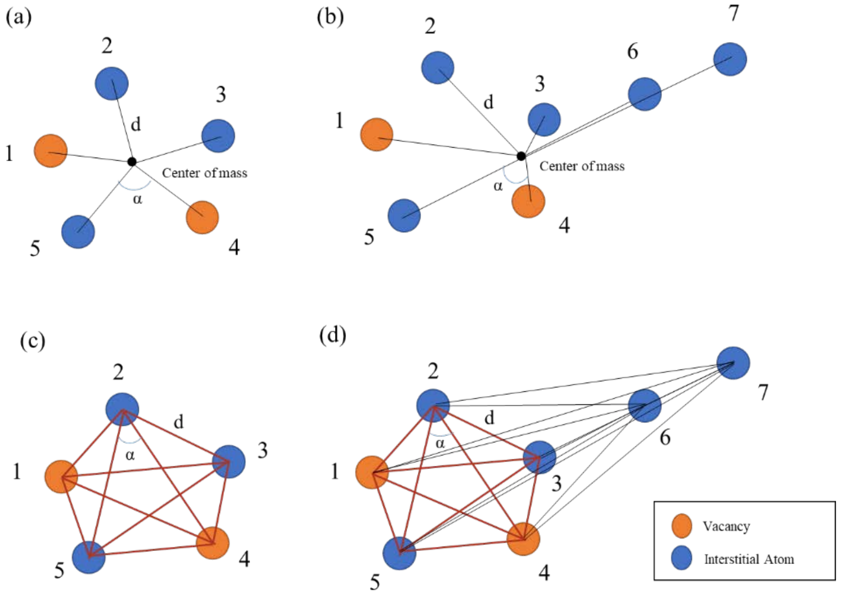

Figure 2 Schematic diagram of the DAPC and DAPP with an example cascade defects

4 Conclusions and Contributions

This paper proposes a novel method and an accurate physical interpretation for feature extraction of cascade defects data, an important particle system in reactor pressure vessel analysis. The comparison of the four feature strategies shows that the DAPP strategy has significant advantages than others in recognizing cascade defects. That is because the center of mass based features may ignore many similar clusters and classify them into different categories. This study shows that feature extraction based on physical information has a positive significance for the analysis of particle systems. This work can be extended not only for clustering of cascade defects, but also for other machine learning studies on particle systems like feature extraction of neural network potential functions which can provide a theoretical support for improving the interpretability of machine learning models on particle systems.

Acknowledgements

This article is the results of the research project funded by National Key R&D Program of China under No.2017YFB0203103 and The National Natural Science Foundation of China under Grant U1867217.

References

[1] M.S. Friedrichs, "Accelerating molecular dynamic simulation on graphics processing units", Journal of computational chemistry, 30.6, 864-872, 2009, https://doi.org/10.1002/jcc.21209.

[2] M.G. Arend, S. Thomas, "Statistical power in two-level models: A tutorial based on Monte Carlo simulation", Psychological methods, 24.1, 1, 2019, https://doi.org/10.1037/met0000195.

[3] C.J. Hu, H. Bai, "Crystal MD: The massively parallel molecular dynamics software for metal with BCC structure", Computer Physics Communications, 211, 73-8, 2017, https://doi.org/10.1016/j.cpc.2016.07.011.

[4] S.G. Li, B.D. Wu, "Massively Scaling the Metal Microscopic Damage Simulation on Sunway TaihuLight Supercomputer", ICPP 2018: Proceedings of the 47th International Conference on Parallel Processing, 1-11, 2018, https://doi.org/10.1145/3225058.3225064.

[5] D.V. Dao, "A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a Monte Carlo simulation", Sustainability, 12.3, 830, 2020, https://doi.org/10.3390/su12030830.

[6] A. Nayarisseri, "Shape-based machine learning models for the potential novel COVID-19 protease inhibitors assisted by molecular dynamics simulation", Current topics in medicinal chemistry, 20.24, 2146-2167, 2020, https://doi.org/10.2174/1568026620666200704135327.

[7] W. L. Jia, "Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning", SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE, 2020, https://doi.org/10.1109/SC41405.2020.00009.

[8] S. Moon, "Automated construction specification review with named entity recognition using natural language processing" Journal of Construction Engineering and Management, 147.1, 2021, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001953.

[9] H. Fujiyoshi, H. Tsubasa, “Deep learning-based image recognition for autonomous driving”, IATSS research, 43.4, 244-252, 2019, https://doi.org/10.1016/j.iatssr.2019.11.008.

[10] F. Özyurt, “Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures”, The Journal of Supercomputing, 1-19, 2019, https://doi.org/10.1007/s11227-019-03106-y.

[11] T. Chen, “Classification with a disordered dopant-atom network in silicon”, Nature, 577.7790, 341-345, 2020, https://doi.org/10.1038/s41586-019-1901-0.

[12] S. Somnath, “Feature extraction via similarity search: application to atom finding and denoising in electron and scanning probe microscopy imaging”, Advanced structural and chemical imaging, 4.1, 1-10, 2018, https://doi.org/10.1186/s40679-018-0052-y.

[13] G.S. Chu, “MD simulation of hundred-billion-metal-atom cascade collision on Sunway Taihulight”, Computer Physics Communications, 269, 108128, 2021, https://doi.org/10.1016/j.cpc.2021.108128.

[14] P. F. Zou, R. F. W. Bader, “A topological definition of a Wigner–Seitz cell and the atomic scattering factor”, Acta Crystallographica Section A: Foundations of Crystallography, 50.6, 714-725, 1994, https://doi.org/10.1107/S0108767394003740.

[15] C. Fiorio, G. Jens, “Two linear time union-find strategies for image processing”, Theoretical Computer Science, 154.2, 165-181, 1996, https://doi.org/10.1016/0304-3975(94)00262-2.

[16] L. McInnes, “hdbscan: Hierarchical density based clustering”, Journal of Open Source Software, 2.11, 205, 2017, https://doi.org/10.21105/joss.00205.

[17] E. Schubert, “DBSCAN revisited, revisited: why and how you should (still) use DBSCAN”, ACM Transactions on Database Systems (TODS), 42.3, 1-21, 2017, https://doi.org/10.1145/3068335.

[18] S. Aranganayagi, “Clustering categorical data using silhouette coefficient as a relocating measure”, International conference on computational intelligence and multimedia applications (ICCIMA 2007), Vol. 2, IEEE, 2007, https://doi.org/10.1109/ICCIMA.2007.328.