Persistent homology analysis with nonnegative matrix factorization for 3D voxel data of iron ore sinters

Ippei Obayashi* and Masao Kimura2,3

1 Center for Artificial Intelligence and Mathematical Data Science (Angels), Okayama University , 3-1-1 Tsushima-naka, Kita-ku, Okayama 700-8530 Japan
2 Photon Factory, Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK) , Tsukuba, Ibaraki 305-0801, Japan
3 Department of Materials Structure Science, School of High Energy Accelerator Science, SOKENDAI (The Graduate University for Advanced Studies) , Tsukuba, Ibaraki 305-0801, Japan

*Corresponding author: i.obayashi@okayama-u.ac.jp

Received

Abstract

This paper proposes a data analysis method using persistent homology and nonnegative matrix factorization. A concatenated persistence image technique is used to extract coexisting structures from the persistence diagrams of different dimensions hidden behind the data. To demonstrate the potential of our method, we apply the method to 3D voxel data of iron ore sinters obtained by X-ray computed tomography. The analysis successfully captures the coexistence structures in these iron ore sinters.

Keywords persistent homology, nonnegative matrix factorization, topological data analysis

Research Activity Group Topological Data Analysis

1. Introduction

Persistent homology (PH) is a data analysis framework that uses topology, which enables us to extract quantitative information about the shape of the data [1]. PH is good at handling disordered and heterogeneous structures and has been applied to various fields, including materials science and life science.

In PH, the data are converted into a two-dimensional (2D) histogram called a persistence diagram (PD) and then analysis is carried out using this diagram. If the PD has some outstanding features, the analysis is easy. In reality, however, we often encounter cases where such features cannot easily be found. One way to deal with this problem is to apply machine learning methods to the PDs. With this method, information about the shape of the data is quantified with PH, and then machine learning is used to find any characteristic patterns.

There are already many studies on the combination of PH and machine learning [2]. There have also been various studies on the vectorization of PDs and kernel methods to use them as inputs for machine learning methods. In this letter, we will present a new method that uses a vectorization method called persistence images (PIs) [3] and a machine learning method called nonnegative matrix factorization (NMF) [4]. By concatenating PI vectors before applying NMF, we can successfully capture the coexistence structures shown in different dimensional PDs. The concatenation technique is simple but works well in combination with NMF.

We also show a case study of analyzing three-dimensional (3D) X-ray computed tomography (X-CT) images of iron ore sinters using the proposed method. Iron ore sinters are the starting material for iron-making processes and is produced by agglomerating fine-size materials, including iron ore particles, limestone flux, and coke breeze through a sintering process. Then, the sinters are reduced in blast furnaces into pig iron [5]. The reduction reaction progresses heterogeneously due to the existence of multiple phases and pore networks in the sinters, and can result in crack formation, one of the main factors in sinter degradation. Characterization of the changes in the distribution of phases during the reduction reactions by the proposed method is expected to provide indispensable information for controlling the microstructures of iron ore sinters.

2. Method

To capture the geometric features of image data, we consider the change of homology in an increasing sequence called a filtration. We often use the signed distance transform to build a filtration from a binary image. Figure 1(b) is the signed distance transform of Fig. 1(a). For each gray pixel, its pixel value is the distance to the nearest white pixel. For each white pixel, its pixel value is the negative distance to the nearest gray pixel. In this paper, we use the Manhattan distance. The filtration Fig. 1(e) is obtained by changing the binarization threshold in Fig. 1(b). In this filtration, homology generators appear and disappear. The theory of PH ensures a unique pairing of the appearance and disappearance of homology generators in the filtration. The pair is called a birth-death pair, and the collection of birth-death pairs is called a PD. A PD is often visualized with a 2D scatter plot or a histogram, as shown in Fig. 1(c)(d).
Here, we utilize the fact that PIs are positive vectors. The discretized vector $\hat{G}$ can regard the result as a vector. The dimension of this vector can be converted to one PD histogram.

$D$ and $w$ are used to reflect the importance of points in $D$ (a birth-death pair near the diagonal is known to be less important than a pair far from the diagonal). One of the advantages of PIs is that vectors of the same dimension can be converted to one PD histogram.

We apply nonnegative matrix factorization (NMF) to the PIs. From nonnegative vectors $\{v_1, \ldots, v_N\}$, NMF finds a small number of nonnegative vectors $\{w_1, \ldots, w_M\}$ and nonnegative numbers $\{\lambda_{n,m}\}_{1 \leq n \leq N, 1 \leq m \leq M}$ that satisfy the following approximation in $\ell^2$-norm:

$$v_n \approx \sum_{m=1}^{M} \lambda_{n,m} w_m.$$  

By applying NMF to PIs, we obtain the following approximation:

$$\rho_{D_k}(X_n) \approx \sum_{m=1}^{M} \lambda_{n,m} \rho_m,$$  

where $\{\lambda_{n,m}\}$ are nonnegative numbers and $\{\rho_m\}_{m=1}^{M}$ are distributions on the $xy$-plane. In this paper, we call $\lambda_{n,m}$ an NMF coefficient and $\rho_m$ a feature distribution.

To integrate PDs of different dimensions, we propose applying NMF to the concatenated vectors. By concatenating three PIs $\hat{\rho}_{D_0}(X_n)$, $\hat{\rho}_{D_1}(X_n)$, and $\hat{\rho}_{D_2}(X_n)$ vertically and applying NMF to the concatenated vectors, we obtain the following approximation:

$$\begin{bmatrix} \rho_{D_0}(X_n) \\ \rho_{D_1}(X_n) \\ \rho_{D_2}(X_n) \end{bmatrix} \approx \sum_{m=1}^{M} \lambda_{n,m} \begin{bmatrix} \rho_{0,m} \\ \rho_{1,m} \\ \rho_{2,m} \end{bmatrix}. \quad (4)$$

In this approximation, we can regard the triple $[\rho_{0,m}, \rho_{1,m}, \rho_{2,m}]$ as a coexisting structure in the PDs of different dimensions. We call $[\rho_{0,m}, \rho_{1,m}, \rho_{2,m}]$ the $m$th concatenated feature distribution and refer to it as CFD$_m$.

By analyzing $\rho_{0,m}$ and $\lambda_{n,m}$, we can understand which birth-death pairs are important for characterizing the PDs. For a deeper investigation of such birth-death pairs, we apply “inverse analysis” to identify the geometric origins of the birth-death pairs [6].

The advantages of our method are summarized as follows:

- NMF is a technique similar to principal component analysis (PCA), and the advantages of NMF compared with PCA are as follows. These are advantageous for the analysis of PDs:
  - Better interpretability because of the positivity of $\rho_m$ and $\lambda_{n,m}$.
  - Each component of NMF tends to reflect the cluster structure of the data [7]. This fact also contributes to the interpretability of the result.
- The technique using concatenated PIs extracts the coexisting structures in the PDs of different dimensions.
- The fundamental idea is simple. Therefore, we can understand and implement the idea easily.

The presented analysis framework is similar to our previous work [8], but the introduction of NMF with concatenated PIs improves the interpretability of the result.

3. Result and discussion

We applied the proposed method to X-CT 3D voxel images of iron ore sinters. We used the same data in previous research [5]. The readers can see detailed information about the data (experimental setup, measurement...
devices, resolutions of the voxel images, etc.) in that paper.

In this experiment, six specimens were measured: two from the early stage of the reduction reaction, two more from the intermediate stage, and the last two from the final stage. The specimens have mainly three phases: iron oxides, calcium ferrites, and large voids. In the preprocessing step, the voxel images corresponding to iron oxides were extracted using an image segmentation technique (the watershed algorithm was used) after applying a Gaussian denoising filter. After binarization, the voxel data were divided into 150 × 150 × 150 cubes to investigate the relative local structures. The cubes corresponding to large initial pores (more than 40% in the cube) were removed since such cubes were considered not to have useful information. PDs were computed from the cubes using HomCloud [1](https://homcloud.dev/).

PIs were also computed using HomCloud, and NMF was performed using scikit-learn [9]. The number of NMF components, \( M \), was set to 3. We tried different vectorization parameters and NMF parameters to check the robustness of the analysis, and we confirmed that the results do not change much even if there are small parameter changes.

By applying inverse analysis to the result, we can identify the geometric origins of \( \rho_{i,j} \). Figure 3 shows some typical examples of the results. \( \rho_{0,0} \) corresponds to branched structures (Fig. 3(a)), \( \rho_{1,0} \) corresponds to tunnels (Fig. 3(b)), and \( \rho_{2,0} \) corresponds to hollow areas (Fig. 3(c)). The NMF results suggest that these three types of structures tend to coexist. Figure 3(d) shows an example of the three types of structures coexisting.

Figure 3(e) shows the schematic picture of the coexisting structures corresponding to CFD\(_0\) as suggested by the obtained results. In the figure, eight ball shapes can be seen very closely together. \( \rho_{0,0} \) corresponds to the center of each ball, \( \rho_{1,0} \) corresponds to the tunnels in the figure, and \( \rho_{2,0} \) corresponds to the cavities surrounded by the balls. From the discussion, we conclude that the tunnel structure of iron oxide with complex branching is considered to correspond to the zeroth concatenated feature distribution. In other words, one typical microstructure of iron ore sinters can be successfully described in terms of the zeroth concatenated feature distribution.

In the same way, the geometric origins corresponding to CFD\(_1\) and CFD\(_2\) can be obtained. It can be considered that CFD\(_1\) (typical of the early stage) corresponds to relatively large grains of iron oxide with bumps (Fig. 3(f)) and that CFD\(_2\) (another structure, typical of the intermediate stage) corresponds to the cluster of small grains of iron oxide (Fig. 3(g)).

The proposed method of the analysis showed that complicated 3D microstructures found in iron ore sinters during the reduction reactions, which previously were not possible to understand quantitatively, can be successfully categorized and quantified. The three typical structures described above can be determined to be the most important shapes that appear during the reduction reactions even without any knowledge of the materials science of the reactions. It should be noted that the determined structures are indeed consistent with our previous researchin terms of the materials science [5]. The structures corresponding to CFD\(_1\) dominate in
Finally, we discuss the limitations of the proposed method. This method works well when there is a one-to-one relationship between the features contained in the PDs of different dimensions. Fortunately, our X-CT data satisfy this condition, but not all datasets do. We consider the following example to illustrate the problem. Suppose the 0-dimensional PDs have two feature distributions $\psi_0,0$ and $\psi_0,1$, the 1-dimensional PDs also have two feature distributions $\psi_1,0$ and $\psi_1,1$, and the 2-dimensional PDs have one feature distribution $\psi_2$. Furthermore, suppose that there is a one-to-one correlation between $\psi_0,0$ and $\psi_1,0$, and between $\psi_0,1$ and $\psi_1,1$, but $\psi_2$ is related to both. In this case, we find the following two concatenated feature distributions using our proposed method:

$$\begin{bmatrix} \psi_0,0 \\ \alpha \psi_1,0 \\ \beta \psi_1,1 \\ \psi_2 \end{bmatrix},$$

where $\alpha, \beta > 0$ are appropriate weights. The interpretation of $\psi_2$ in this case is more difficult than the example analysis of iron ore sinters. In other words, if the data contain more complex correlations between features, the advantage of our method, which is good interpretability, is lost.

**Acknowledgments**

This work is partially supported by JSPS KAKENHI JP 19H00834, JSPS Grants-in-Aid for Transformative Research Areas (A) 20H05884 and 22H05109, and JST PRESTO JPMJPR1923.

**References**

[1] I. Obayashi, T. Nakamura and Y. Hiraoka, Persistent Homology Analysis for Materials Research and Persistent Homology Software: HomCloud, J. Phys. Soc. Jpn., 91 (2022), 091013.

[2] C. S. Pun, S. X. Lee, and K. Xia, Persistent-homology-based machine learning: a survey and a comparative study, Artif. Intell. Rev. (2022).

[3] H. Adams et al., Persistence Images: A Stable Vector Representation of Persistent Homology, J. Mach. Learn. Res., 18 (2017), 1–35.

[4] D. Lee and S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 401 (1999), 788–791.

[5] M. Kimura et al., Non-empirical identification of trigger sites in heterogeneous processes using persistent homology, Sci. Rep., 8 (2018), 3553.

[6] I. Obayashi, Volume Optimal Cycle: Tightest representative cycle of a generator in persistent homology, SIAM J. Appl. Alg. and Geom., 2 (2018), 508–534.

[7] C. Ding, X. He, and H. D. Simon, On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering, in: Proc. of the SDM’2005, pp. 606–610, SIAM, 2005.

[8] I. Obayashi, Y. Hiraoka and M. Kimura, Persistence diagrams with linear machine learning models, J. Appl. and Comp. Topol., 1 (2018), 421–449.

[9] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12 (2011), 2825–2830.