パーシステントホモロジーの理論と応用

大林一平 [1,2]

[1] 理化学研究所 革新知能統合研究センター
[2] 東北大学 材料科学高等研究所

概要

本講演ではパーシステントホモロジーの理論から応用、ソフトウェアまでを解説する。

Theory and applications of persistent homology

Ippei Obayashi[1,2]

[1] Center for Advanced Intelligence Project, RIKEN
[2] Advanced Institute for Materials Research, Tohoku University

Abstract

In this presentation, I introduce the theory, applications, and software for persistent homology.

1 はじめに

図 1 の左側上下の画像は MD シミュレーションで計算した液体シリカとアモルファスシリカの原子配置であるが、この 2 つの特徴的な違いは何だろうか? またこのデータのどの部分の構造を見ればその特徴がわかるであろうか?

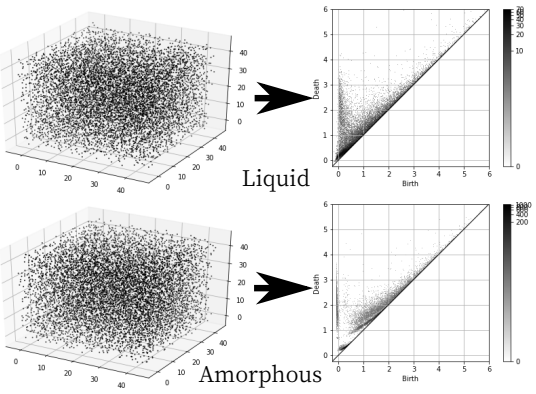

図 1: 液体シリカとアモルファスシリカの原子配置およびそのパーシステント図。液体シリカはパーシステント図は点の分布がブロードだが、アモルファスシリカのパーシステント図は筋状に点が集中して分布している。[1] のデータより再構成したもの。

こういった問題を解決するためには、データから形状の情報を何らかの形で定量的に抽出する必要がある。パーシステントホモロジーはこういった問題を解決するためのツールとして利用できる。図 1 の右側上下の画像はそれぞれの原子配置データから計算された 1 次のパーシステント図である。このようにパーシステント図を見るとこの 2 つには何か明らかな違いがあることがわかる。

本講演では、このパーシステント図とは何か、どのようにこの図を解釈すれば良いのか、どうやったら計算できるか、などについて解説する。

2 パーシステントホモロジー

位相的データ解析という、数学のトポロジーの概念を使ったデータ解析手法がここ 20 年、理論やソフトウェアから応用まで急速に発展しつつある。パーシステントホモロジー [2, 3] は位相的データ解析の中でも特に重要な概念である。パーシステントホモロジーは図形の孔、空隙、連結成分、といった構造に注目することによって、データの形の情報を定量的かつ効率的に抽出することができる。パーシステントホモロジーは材料科学 [1, 4, 5, 6] や生命科学 [7, 8] 等様々な分野でのデータ解析に応用されている。

2.1 パーシステントホモロジーの定義

ではパーシステントホモロジーについて説明していこう。図 2(a) の点集合データから何らかの形の情報を得たい、とする。このデータには当然孔はないが、孔のようなものがあるように見える。そこで (b) のように半径 r の円を各点に置くことで孔を作りだす。この図の場合では大中小 3 つの孔が現われる。

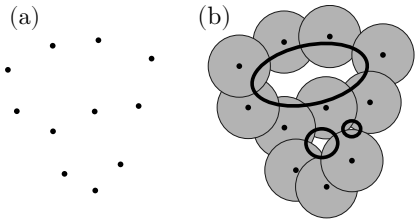

図 2: (a) 点集合データ (b) その上に円を置いたもの

しかし、ここには 2 つの問題がある。適切な半径の決め方の問題と、孔の大きさの情報がない (個数しかわからない) という問題である。ではどうするのかというと、増大列 (数学的にはフィルトレーションと呼ばれる) を使うのである。図 3 のように各点に置いた円の半径を徐々に大きくしていくことを考える。すると最初は孔がなかったのが半径 r_2 で孔が生じ、半径 r_3 でその孔が 2 つに分かれ、r_4 でその一方が消滅し、最後に r_5 でもう一方の孔が消滅する。パーシステントホモロジーの理論によってこの孔の生成と消滅をうまく対応付けることができる。この場合には r_2 で現れた孔が r_5 で消え、r_3 で現れた孔が r_4 で消える、と対応付けることができる。このとき、$(r_2, r_5), (r_3, r_4)$ というペアリングを birth-death pair と呼び、ペアの各要素を birth time、death time と呼ぶ。これらの birth-death pair の集まりをパーシステント図 (Persistence diagram, PD) と呼ぶ。PD の可視化のためには、これを平面上にプロットしたり、ヒストグラムを描いたりする。

PD はデータの形の情報をうまく縮約していると考えられ、実際様々なデータ解析に利用されている。birth time、death time の意味付けを考えると、birth time は孔をぐるっと囲むリング状の点の点間距離の 1/2 の最大値、death time はそのリングの部分に入る円の半径の最大値、ということがわかる。つまり birth time と death time には孔を囲う点の密度と孔の大きさが符号化されていることがわかる。また birth time と death time の差は発生した孔がどれほど長持ちするかを意味する。つまりこの差 (生存時間と呼ばれる) は点の重要度を決める要素として利用できる。

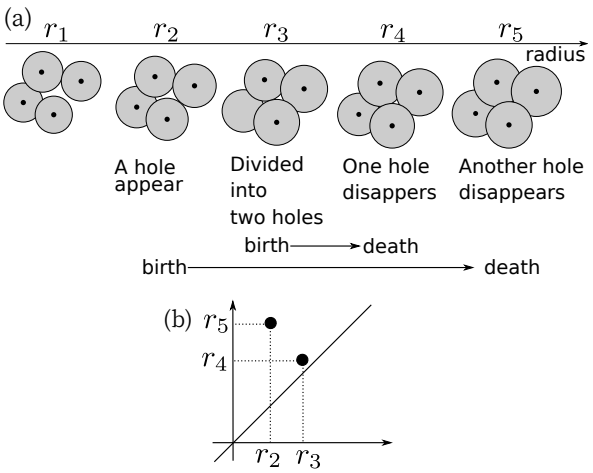

図 3: パーシステントホモロジーの例 (a) 増大列の例 (b) 対応するパーシステント図

孔の種類には実は次元のようなものがあり、ホモロジー理論で次数と呼ばれる。そのため PD にも次数がある。1 次の PD は「孔」「リング構造」に注目していて、2 次は「空洞」に注目している。0 次の PD もあり、実はこれは孔ではなく連結成分に注目している。

2.2 シリカの原子配置の例

それでは本稿冒頭に紹介したアモルファスシリカと液体シリカの例についてついて解説しよう。アモルファスシリカは共有結合によるネットワーク構造が特徴的で、そのネットワークがなすリング構造に重要な特徴がある。そこで 1 次の PD が有用である。

図 1 の右側の 2 つの図がそれぞれの原子配置から計算された 1 次元のパーシステテント図である。見てわかるように、アモルファスシリカの PD にはいくつか筋状の分布が見られる。これはアモルファスシリカの原子配置にはある種の典型的な「形」があることを示唆している。そしてこれは実は隣接していない原子の配置にある制約があることを意味している。そしてこの「隣接していない原子の配置の制約」がアモルファスの中距離秩序と対応していると考えられる。

3 応用

この節ではパーシステントホモロジーを使った解析についていくつか紹介する。

3.1　ポリマーの伸長変形の解析

まずはポリマーの伸長変形による破断現象の解析 [5] について紹介する。この研究ではポリマーをビーズ状に粗視的にモデル化し、MD シミュレーションでその挙動を計算した。シミュレーション内で 1 軸方向に伸長変形することで、ポリマーに小さなひびが入り、それが成長し、最終的には塑性変形[1]する。

このひびの成長を調べたい、というのがこの研究の目標である。しかしシミュレーション内ではポリマーはビーズとしてモデル化しているので、ひびの定義自体が自明でなく、それを特定することは簡単ではない。これがパーシステントホモロジーの役割である。ここでは 2 次の PD に対応する空隙がひびである、とみなして解析している。

この研究で有効な道具が PD の逆解析と呼ばれるものである。データから PD を計算するのは 2 節で説明した方法でできるが、逆に PD の上の birth-death pair がどのようなリング構造、空隙構造に対応しているのかを特定するのはまた別の問題となる。これは逆解析と呼んでいる問題の一例で、最近研究が進んでいる [9, 10, 11]。これによってひびの具体的な形を特定することができる。

この論文の最終的な結論としては小さなひびが数多く繋がっていくことで大きいひびに成長するという「ひびのパーコレーション」とでも呼ぶべき現象が見られることを示した。

3.2　機械学習との組み合わせ

パーシステントホモロジーと機械学習/統計の組み合わせは非常に強力なデータ解析手段である。パーシステントホモロジーによってデータの形の情報を抽出し、機械学習によってそのパターンを認識する。つまりこの組み合わせによってデータの背後に隠された特徴的幾何パターンを取り出すのである。

本講演では [12] で提案された枠組みについて解説する。この手法は焼結鉱の X 線 CT 画像の解析 [6] に実際に利用されている。

図 4 は全体の枠組みを示している。PD をベクトルに変換して機械学習の入力とすることは機械学習の基本的アイデアである。機械学習の用語で言うと、PD を特徴量として利用しているのである。

さらにこれを上で述べた PD の逆解析と組み合わせることで、学習で得たパターンを元データの上にマッピングすることができる。これによってデータの背後にある幾何的パターンを直感的に理解することが可能となる。

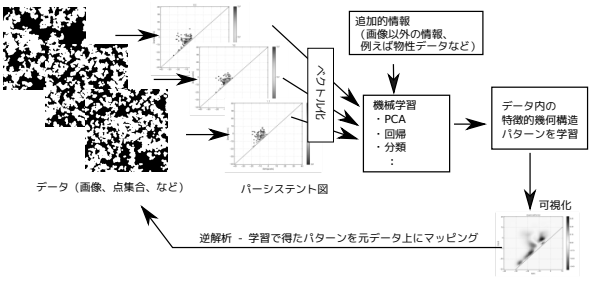

図 4: 機械学習の全体像

4　ソフトウェア

パーシステントホモロジーの計算ソフトウェアは数多くある。名前だけ紹介すると、Gudhi, dipha, phat, ripser, eirine, RIVET, JavaPlex, Perseus, Dionysus, などである。それぞれのソフトウェアはそれぞれの開発者がそれぞれの興味関心にもとづいて開発している。本講演では特に我々が開発している HomCloud[2]について紹介する。

HomCloud の大きな特徴は、応用、特に材料科学への応用、を重視していることである。具体的な利用事例にフォーカスし、可視化、逆解析、統計との連携、といった部分を重視して開発している。パーシステントホモロジーの計算自体は既存のソフトウェア (上で挙げた dipha と phat) を利用してる。利用可能なデータは 2/3 次元の点集合データ、もしくは n 次元のピクセルデータ、である。

HomCloud は現在上に挙げた URL から自由にダウンロードして利用できる。インストールの仕方や基本的なデータ解析のためのチュートリアルなども準備してあるのでそちらも参考にしてほしい。

HomCloud は現在も活発に開発中で、1、2 ヶ月に 1 度の頻度でバージョンアップされている。バージョンアップごとにバグの修正、ドキュメントの改善、新機能の追加、と様々な改良が随時加えられている。

[1] 加えている力を解放しても元に戻らない変形。対して弾性変形は力を加えなくすると元に戻る。

[2] https://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud/index.html

5　おわりに

パーシステントホモロジーは形の情報を定量化する、これまでにない新しいツールである。原子配置のような点集合データ、また 2 次元/3 次元画像データの解析に利用可能である。特にアモルファスや多孔質といったような非一様的で乱れた、しかし完全にランダムでもないデータへの応用に有効である。機械学習との併用も有効で、これによってデータの背後にある特徴的幾何パターンを取り出すことができる。これらの解析のためのソフトウェアも整備されつつある。パーシステントホモロジーによるデータ解析には難しい点も多いが、これまでにない結果を得ることができる可能性のあるツールである。

参考文献

[1] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G. Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016.

[2] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

[3] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, Feb 2005.

[4] M. Saadatfar, H. Takeuchi, V. Robins, N. Francois, and Y. Hiraoka. Pore configuration landscape of granular crystallization. *Nature Communications*, 8:15082, 2017.

[5] Takashi Ichinomiya, Ippei Obayashi, and Yasuaki Hiraoka. Persistent homology analysis of craze formation. *Phys. Rev. E*, 95:012504, Jan 2017.

[6] Masao Kimura, Ippei Obayashi, Yasuo Takeichi, Reiko Murao, and Yasuaki Hiraoka. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. *Scientific Reports*, 8:3553, 2018.

[7] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[8] Cang Zixuan, Mu Lin, Wu Kedi, Opron Kristopher, Xia Kelin, and Wei Guo-Wei. *cmb*, volume 3, chapter A topological approach for protein classification. 2018 2015. 1.

[9] Emerson G. Escolar and Yasuaki Hiraoka. *Optimal Cycles for Persistent Homology Via Linear Programming*, pages 79–96. Springer Japan, Tokyo, 2016.

[10] B. Schweinhart. *Statistical Topology of Embedded Graphs*. PhD thesis, Princeton University, August 2015. `https://web.math.princeton.edu/~bschwein/`.

[11] Ippei Obayashi. Volume optimal cycle: Tightest representative cycle of a generator in persistent homology. *SIAM Journal of Applied Algebra and Geometry*, 2018. Accepted. `https://arxiv.org/abs/1712.05103`.

[12] Ippei Obayashi, Yasuaki Hiraoka, and Masao Kimura. Persistence diagrams with linear machine learning models. *Journal of Applied and Computational Topology*, 1(3):421–449, Jun 2018.