パーシステントホモロジーに基づくデータ解析パッケージ HomCloud の紹介

HomCloud, data analysis package based on persistent homology

理化学研究所・革新知能統合研究センター大林一平
Ippei Obayashi
Center for Advanced Intelligence Project, RIKEN

Abstract

本稿では筆者が中心として開発を進めているパーシステントホモロジーにもとづくデータ解析ソフトウェア HomCloud の 2020 年 5 月における現状とその将来の開発予定について紹介する．

1 はじめに

パーシステントホモロジー (PH) は数学のホモロジーの概念をデータ解析に利用するための数学的枠組みである [1, 2]．PH を利用することでデータの形の情報を定量的に抽出することが可能となる．材料科学 [3, 4, 5]，分子遺伝学 [6]，生化学 [7] など様々な領域での応用が進みつつある．

数学的には PH はフィルトレーション (位相空間の増大列) 上のホモロジー理論で，フィルトレーションにスケールなど連続的な情報をエンコードすることでホモロジーの情報にそういった連続的情報を付加することを可能とする．

具体的に簡単なグレイスケール画像を用いてアイデアを説明しよう．ここで考えるのは図 1(a) のようなデータである．これは 5x3 ピクセルのグレイスケール画像で，グリッドの中の数字がそのグレイスケールのレベルである．ここからホモロジーを考えるためには適当なっ閾値で画像を二値化する必要がある．例えば 4 以下で二値化すると図 1(c) のようになり，このホモロジー群は $H_0 \simeq \mathbb{Z}, H_1 \simeq \mathbb{Z}$ となる．一方 5 以下で二値化すると図 (d) のようになり，このホモロジー群は $H_0 \simeq \mathbb{Z}, H_1 = \mathbb{Z}^2$ となる．すると閾値の決め方で得られる結果が当然変わる．ここで閾値を図 1(b) から (f) のように変えていったときのホモロジーの変化がこのデータのトポロジカルな情報と考えられる．ではこれをうまく捕捉する方法はないだろうか？

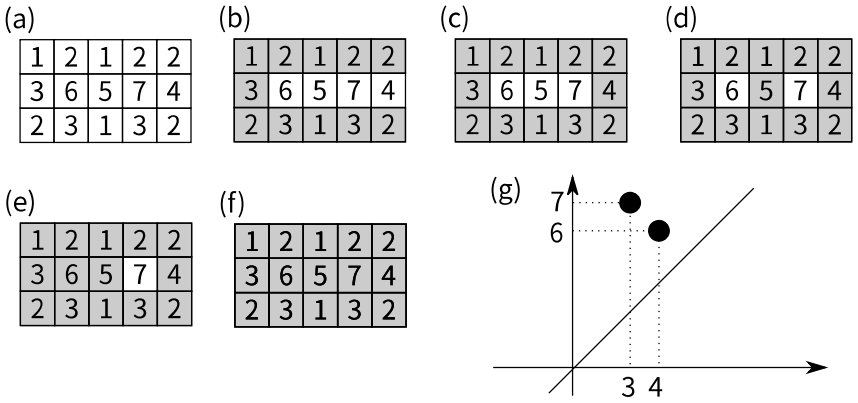

Figure 1: グレイスケール画像のパーシステンス (a) グレイスケール画像 (b) (レベル) ≤ 3 での二値化 (c) (レベル) ≤ 4 (d) (レベル) ≤ 5 (e) (レベル) ≤ 6 (f) (レベル) ≤ 7 (g) 1 次のパーシステント図

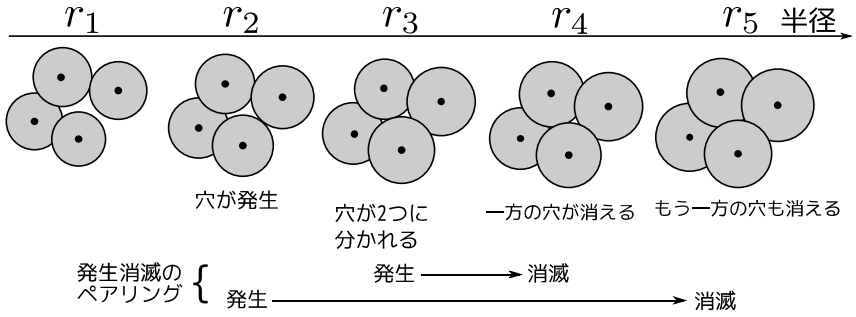

Figure 2: ポイントクラウドの PD の計算

一つの方法としては閾値を横軸，ベッチ数 (連結成分の数，穴の数) を縦軸にしてグラフを描くことだろう．しかしもっとうまい方法がある．実は図 1での穴 (つまり 1 次のホモロジー) の発生消滅の過程から個々の穴の発生と消滅のペアを取ることができるのである．この図では閾値 4 で発生した穴が閾値 7 で消滅し，閾値 5 で発生した穴が閾値 6 で消滅する，と対応する．これはホモロジー準同型写像が生成元をどのように移すかを調べればわかる．実は PH の構造定理と呼ばれる定理によって，ホモロジーの係数に体を使う場合，つまりベクトル空間を考える場合にはこのようなペアリングが常に一意に可能であることがわかる．この発生と消滅の閾値のペアを図 1(g) のように XY 平面にプロットしたものをパーシステント図 *(Persistence diagram, PD)* と呼び，各ペアを *birth-death pair* と呼ぶ．

グレイスケール画像の他にも様々なデータに対して PH は利用できる．例えばポイントクラウド (有限個の点の集合) の場合には各点に円盤 (2 次元) や球 (3 次元) を置き，その半径を大きくしていくことでフィルトレーションを構成する (図 2)．白黒画像の場合は distance transform という距離を使った変換 (数学的には距離関数を使う) でグレイスケール画像化してから PH を適用する．それ以外にも距離行列 (つまり n 個の点があり，その各ペアの間で距離が決まっているもの) に対しては Vietoris-Rips 複体と呼ばれる複体を使うことでフィルトレーションを構成する．単体複体や胞体複体の増大列が構成できるならばなんでも PH の対象とできる．

例えば，[3] では分子動力学シミュレーションで計算したアモルファスの原子配置の解析を行っているが，これは原子の中心を点としたポイントクラウドとして球を膨らませていく方針で解析を行っている．[6] ではウィルスの個々の遺伝情報を点をみなし，その間の距離を遺伝的な距離で定義することで Vietoris-Rips 複体を使っている．[4, 5] などでは二値画像データに signed distance transform[1]を使って解析を行っている．

1.1 PH のソフトウェア

このように PH の応用は徐々に進みつつあるが，応用に重要なのはデータ解析のためのソフトウェアである．PH のためのソフトウェアには例えば以下のようなものが挙げられる: Gudhi[2], dipha[3], phat[4], ripser[5], eirine[6], RIVET[7], JavaPlex[8], Perseus[9], Dionysus[10]．これらのソフトウェアはそれぞれの開発者の研究の方向性や応用分野に応じて開発されている．例えば dipha, phat, ripser は同一の研究グループで開発されているが，主な関心は PD の計算アルゴリズムで実際優れた性能を持っている．Gudhi は複体の多様な表現にフォーカスしている．

本稿では以下筆者の大林が中心となって開発している HomCloud について解説する．2020 年 5 月における HomCloud の現状と将来計画について述べる．

[1] distance transform に正負の符号をつけたもの
[2] https://gudhi.inria.fr/
[3] https://github.com/DIPHA/dipha
[4] https://bitbucket.org/phat-code/phat/
[5] https://github.com/Ripser/ripser
[6] http://gregoryhenselman.org/eirene/
[7] https://github.com/rivetTDA/rivet
[8] http://appliedtopology.github.io/javaplex/
[9] http://people.maths.ox.ac.uk/nanda/perseus/
[10] https://www.mrzv.org/software/dionysus/

2 HomCloud について

HomCloud は筆者の大林が中心となって開発している PH にもとづくデータ解析ソフトウェアである．ウェブサイトの URL は`https://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud/` である．フリーソフトウェアライセンスで誰でも自由に利用，改造することが可能である．

HomCloud は大林が東北大の材料科学高等研究所に所属していたころに開発を始めたという経歴があり，材料科学への応用を当面の目標として開発を開始した．そのころの対象となるデータとしては分子動力学シミュレーションによる原子配置データ (3 次元ポイントクラウドとして扱う)，各種顕微鏡による画像 (2D が多いが，X 線 CT などを使った 3D 画像なども解析対象) などがある．現在はより一般的なデータ解析の応用を目指し，

- 2,3 次元のポイントクラウド
- n 次元の画像 (ピクセルデータ，ボクセルデータ，など)
- 距離行列
- 抽象有限複体 (つまり境界作用素を明示的に与える)

などが入力データとして利用可能である．3 次元ポイントクラウドデータは各点に初期半径を設定することが可能で，これは各原子固有の半径を表現するためなどに使われる．利用頻度が高いのはポイントクラウド，2D/3D の画像データで，これらのデータに対する機能は他のデータのよりも充実している．

HomCloud は可視化，機械学習，逆解析といった応用的な機能にフォーカスして開発を進めている．元々が材料科学への応用が目的であり，現在でも材料科学に限らない応用を第一目標として開発を行っている．PD 計算そのものについては dipha, phat, ripser などが強力であるため，これらをバックエンドで利用している．

特に逆解析機能は HomCloud が最も先進的な機能を持っている．これは簡単に言うと PD の各点に対応する幾何構造を特定するための機能である．HomCloud は Optimal Volume[8] という概念をベースとした高度な逆解析機能を持っている．現状では 2,3 次元ポイントクラウドと n 次元の画像にしか対応していないが，良く使われるこれらのデータの解析に対して逆解析が可能なのは大きな長所と言える．PD が捉えている構造を元データの上にマッピングすることで PD の計算結果を直感的に理解すること可能とする．マッピング結果からさらなるデータ解析も可能である．

機械学習については Persistence Image(PI)[9, 10] という手法を用いた PD のベクトル化機能が利用可能である．PI は PD のヒストグラムを高さ方向に考えてベクトルを見なすベクトル化の手法で，単純で理解しやすい割には学習性能も良い手法である．機械学習や統計学の手法では同じ次元のベクトルデータを入力として必要とするので，PI で変換したベクトル値を使うのである．PI はベクトルを逆にヒストグラムに変換することも可能で，それが学習結果の解釈性に寄与している．そういった単純さと性能のバランスを考慮して HomCloud では PI を推奨している．

可視化周辺は PD を見た目よく出力するだけでなく，逆解析の結果の可視化なども考慮して実装されている．2 次元，3 次元データの PH 解析などにこういった逆解析結果の可視化は有用である．

HomCloud は主に Python によって実装されている．一部性能が要求される部分だけは C++ で記述されている．これは Python の科学技術計算エコシステムを活用し，HomCloud を効率的に実装するためである．同時に HomCloud の利用者がこういったエコシステムと組み合わせてデータ解析をできるようにするためでもある．ユーザインターフェースとしてはコマンドラインと Python インターフェースの 2 通りが利用可能である．特に Python インターフェースは Numpy の配列を入出力に使うので Python のエコシステムと容易にやりとりできる．

HomCloud の開発方針として一つ強調しておきたいこととしては，開発者 (つまり本稿の筆者) がデータ解析とソフトウェア開発，そして理論研究を平行して行っていることである．開発者が利用者を兼ねることで「本当に必要な機能は何か」「この機能は使いやすいか」といった観点を開発に効率的に取り入れることができる．さらに理論研究を並列してやることで新しい手法を迅速に HomCloud に取り入れることができる．

HomCloud は現在 Linux, Mac, Windows をサポートしている．筆者は Linux を普段使っているため Linux が一番確実に動くが他の環境でも動作する．機能はどの OS でも同じである．各 OS ごとのインストールガイドは`https://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud/install-guide/`

index.html から参照することができる．HomCloud を利用したい場合はこちらを参考にインストールを進めると良い．

インストール後はチュートリアル (https://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud/basic-usage.html) を順にやっていくとよい．コマンドライン版と Python 版が用意されている．ポイントクラウドデータ解析のチュートリアルから始めて，その後は興味のあるチュートリアルをするとよいだろう．ポイントクラウド解析の他にも，画像 (2D 二値画像，2D グレイスケール画像，3D 二値画像)，ポイントクラウドの機械学習などのチュートリアルが利用可能である．

このチュートリアルを済ませればとりあえず PH によるデータ解析は可能になるだろう．ただ，データの前処理の問題，解析結果の解釈の問題，解析手法の選択の問題 (機械学習が使えるのかどうかなど) など様々な問題が実際的には生じると思う．共同研究の相談は常に歓迎するので，そういった場合には我々に相談してほしい．

3 HomCloud の将来について

HomCloud は現在各種研究費のバックアップを受けて改良を進めている．現状では以下のような改良，機能追加などを予定している．

- Windows や Mac のサポートの強化
- コードベースの改善
- 3 次元画像解析の改善
- 機械学習関連の機能の強化
- 距離行列関連の機能強化
- その他

筆者は Linux を使っているため，それ以外の Mac や Windows のサポートは一歩遅れている．Mac については筆者が研究費で雇っているアルバイトの人に使ってもらっているためましだが，Windows についてはさらにもう一歩階遅れている[11]．こういった問題を解決するためにテストやパッケージビルドの自動化を行い，問題を早期に発見できる仕組みを作る予定である．

HomCloud は 3 年以上開発を進めているため，徐々に過去のコードが改良の妨げになりつつある状況にある．そこでコードベースの改善を行い，今後の改良や機能追加に備える予定である．この改良は上で述べた Windows や Mac でのテスト自動化のためにも必要なので，2020 年度内に自動化も含め完了したいと考えている．

3 次元画像解析は最近筆者が関わっているプロジェクトでも必要となっているので改良を進めている．現状は PD 計算の高速化について改良を進めている状況である．その他にも可視化関連，特に逆解析結果の可視化についてはより利便なツールが欲しいと考えている．ボクセルデータ上の逆解析結果を観察するためには 3 次元データを画面上で回転させるだけでは困難を感じてきたため，より良い仕組みを考えてたい．ここについては具体的なアイデアがあるわけではないので利用者からの意見も欲しいと考えている．

機械学習については既に PI でのベクトル化を実現していることを説明したが，他にも機械学習のための様々なベクトル化の手法が提案されている．手法によって長所短所があるが，それらを比較するためにも様々な手法をサポートしたいと考えている．

「はじめに」で説明した通り，距離行列を入力にする場合には Vietoris-Rips 複体という複体を利用して PD を計算する．今まではあまり利用者がいなかったため関連の機能は少なかった[12]．例えば逆解析機能はポイントクラウドなどと比べても貧弱であった．しかし最近は筆者が関連しているプロジェクトでの利用が始まったため，この周辺の機能強化を図っていく予定である．そのプロジェクトからの要望が強いため，優先順位を高くして進めていく．

それ以外にも研究成果を随時 HomCloud に追加する予定である．最近プレプリントを出した結果 [11] も実は HomCloud に実装済みである．この論文は理論寄りの結果で実は実用性はそれほど高くはないのだが，こういった理論研究の成果も HomCloud に搭載していく．

[11] ただ Mac 環境は OS や homebrew などのバージョンアップで問題が生じやすい傾向にあると感じており，そのあたりは Windows のほうがましだと考えている．

[12] これは開発ポリシーとも関係あり，利用者のない機能を強化するのはソフトウェアの不要な肥大化を生むため避けてきた．

4 おわりに

本稿ではパーシステントホモロジーに基づくデータ解析パッケージ HomCloud についてその現状について紹介し，将来の予定について述べた．HomCloud は理論と応用を繋ぐ架け橋として重要であると筆者は考えており，今後も発展させていく予定である．また読者の利用者からのフィードバックも歓迎する．バグ報告や要望などをお願いしたい．より積極的な貢献 (例えば不足しているドキュメントの提供や紹介記事の執筆など) があればさらにありがたい．単に利用するだけでも HomCloud に対する貢献である．HomCloud を利用した論文を出すときはソフトウェアの URL に加えて本稿を参考文献に加えると良いだろう．

謝辞

HomCloud の開発には，JSPS 科研費 JP 16K17638 および JP 19H00834，JST さきがけ JPMJPR1923，JST 未来社会創造事業 JPMJMI18G3，戦略的イノベーション創造プログラム（SIP）統合型材料開発システムによるマテリアル革命，の助成を一部受けている．

References

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

[2] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, August 2017.

[3] 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.

[4] Vanessa Robins, Mohammad Saadatfar, Olaf Delgado-Friedrichs, and Adrian P. Sheppard. Percolating length scales from topological persistence analysis of micro-ct images of porous materials. *Water Resources Research*, 52(1):315–329, 2016.

[5] 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(1):3553, 2018.

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

[7] Cang Zixuan, Mu Lin, Wu Kedi, Opron Kristopher, Xia Kelin, and Wei Guo-Wei. A topological approach for protein classification, 2015.

[8] I. Obayashi. Volume-optimal cycle: Tightest representative cycle of a generator in persistent homology. *SIAM Journal on Applied Algebra and Geometry*, 2(4):508–534, 2018.

[9] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

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

[11] Ippei Obayashi and Michio Yoshiwaki. Field choice problem in persistent homology. arXiv:1911.11350, 2019.