The theory and applications of persistent homology

Ippei Obayashi

Center for Advanced Intelligence Project (AIP), RIKEN
Advanced Institute for Materials Research (AIMR), Tohoku University

Nov. 5, 2018
Outline

1. Introduction
2. Homology and persistent homology
3. Applications of persistent homology
4. Software for persistent homology
Introduction
Persistent homology

- **Topological Data Analysis (TDA)**
  - Data analysis using topology from mathematics
  - Characterize the shape of data quantitatively
    - Connected components (islands), rings (holes), cavities

- **Persistent homology (PH) is one of the most important tools for TDA**
  - Uses the concept of “homology”
  - Gives the good descriptor of the shape of data (persistence diagram)

- **Developed rapidly in 21st century**
  - Mathematical theories and algorithms
  - Software
  - Applications to materials science, life science, etc.
Mathematics and data analysis
  ▶ Probability - statistics and machine learning
  ▶ Analysis - Fourier analysis and numerical analysis
  ▶ Algebra - Symmetry analysis (for crystals)
  ▶ Geometry and topology - TDA

TDA is good for:
  ▶ heterogeneous data
  ▶ disordered data
  ▶ data without complete randomness

Mathematics and materials
  ▶ Liquid and gas - random - probability theory and statistical models
  ▶ Crystals - ordered - group theory
  ▶ Amorphous, polycrystalline, and porous media - disordered - topology
Example 1

Atomic configurations of amorphous silica and liquid silica. Do you identify?
We can identify by using persistence diagram.
Example 2

What is the characteristic difference between these two pointclouds?
We can distill the characteristic geometric patterns by the combination of PH and machine learning.
Homology and Persistent homology
Homology

- We can mathematically formalize “connected components”, “rings” “cavities” by homology.
- Algebra is used for the formalization
- We can identify the “type” of “holes” by a kind of dimension (called degree)

1 dim: You can see the inside from outside
2 dim: You cannot see

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Shape 1</th>
<th>Shape 2</th>
<th>Shape 3</th>
<th>Shape 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dim 1:</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dim 2:</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Count the rings

How many rings in this figure?
Linear algebra is the key to count the rings. Here we have \((1) + (2) + (3) = (4)\) since two arrows with opposite directions are canceled. Therefore these four rings are \textit{linearly dependent}, and we can count the number of \textit{linearly independent} rings by using linear algebra.
Persistent homology

- Characterize the shape of data is difficult problem
  - for 3D data or higher dimensional data.
- Homology is used for that purpose, but we can only count the number of holes
- We need better way than homology
- Computational homology is not robust to noise.

→ Use increasing sequences (filtrations)
Input data is a set of point (a pointcloud)
There is no holes in this pointcloud, but it looks like some holes
Put discs of radii $r$ on all points
Three holes
  - We can count the holes by homology
Filtration

As the radius $r$ become larger, some holes appear and disappear. We can make pairs of appearance and disappearance of a hole by using mathematical theory of PH.

I. Obayashi (AIP, Riken)

Theory and applications of PH

Nov. 5, 2018 17 / 38
These pairs are called *birth-death pairs*. and the set of all birth-death pairs are called *persistence diagram* (PD).

**1st persistence diagram**
PH is applicable to any dimensional data
  ▶ But it is hard to intuitively understand higher dimensional holes, 2D or 3D data is easy to analyze
  ▶ Especially, PH is useful for 3D data

Various increasing sequence
  ▶ Image data
  ▶ Especially 3D data, such as X-ray CT scan data
The following two mathematical theorems are important:

- **Structural theorem for PH**
  - Gives an algorithm of PDs
  - Uniqueness of a PD for a given input data

- **Stability theorem for PH**
  - Ensures the robustness of a PD to noises
Applications
Craze formation of polymers

Kremer-Grest model

- uniaxial deformation

void coalescence during craze formation

- gray voids are large voids observed after yielding
- color voids are initial micro voids generating large voids

- detect large voids from PD movie by generators with large death values
- explore initial configurations of large voids by reversing time
- large voids are generated by coalesce of micro voids (void percolation)
The atomic configuration of amorphous silica looks like random
  Similar to liquid silica
But amorphous silica has rigidity.
Some geometric structures are important for the rigidity.
シリカの原子配置

識別可能？ or 両者ともにランダム？
シリカのパーシステント図

- PD1を表示（リング構造に着目）
- 結晶の規則性は0次元的分布
- 液体のランダム性は2次元的分布
- ガラスは1次元的分布（曲線）!!
ガラスの階層的幾何構造

階層的リング構造

- Cp: primary rings generating the others
- Ct: triangles on tetrahedra
- Co: three oxygen rings
- Bo: oxygen rings (≥ four)

逆問題
- optimal cycle
  Escolar and H. 2015.
- continuation
  Gameiro, Obayashi, H. Physica D, 2015

PD内の曲線の幾何学的な起源
Combination of Machine learning (ML) and PH

We have 200 pointclouds

- 100 pointclouds are labeled by 0, and other 100 pointclouds are labeled by 1
- Find characteristic geometric patterns by ML and PH
Framework

- Data (point clouds, images, etc.)
- Persistence diagrams
- Machine learning
  - PCA
  - Regression
  - Classification
- Characteristic geometric patterns in data
- Additional information
- Inverse analysis
- Visualize
Each pointcloud is transformed into a PD
- Vectorize PDs and apply a machine learning method
- We can visualize the learned result in the form of a PD
- We can identify important birth-death pairs by comparing the learned result.
- The important pairs are mapped on the original input data by using the “inverse analysis of PDs”
- Please see the demo
Software
Software

Software is important for practical data analysis by PH. I introduce you *HomCloud*, data analysis software based on PH.
Various software

There are many software for PH.

- Gudhi
- dipha, phat, ripser
- eirine
- RIVET
- JavaPlex
- Perseus
- Dionysus
HomCloud

- Focus on applications, especially to materials science
  - MD simulation data
  - 2D/3D image data
  - Easy installation, user interface, machine learning, inverse analysis
We can compute PDs from 2D/3D pointclouds and $N$ dimensional bitmap data.
逆解析
HomCloud Demo
Summary

- We can analyze the shape of data effectively and quantitatively by using PH
  - Based on topology
  - PDs are good descriptors for the shape of data
  - Useful for 3D data

- Various applications
  - Materials science
  - Life science, geology, etc.

- The fusion of theoretical studies, software development, and practical data analysis is important.
Appendix