Mapper and its Applications

Piyush Agarwal

IIT Bombay

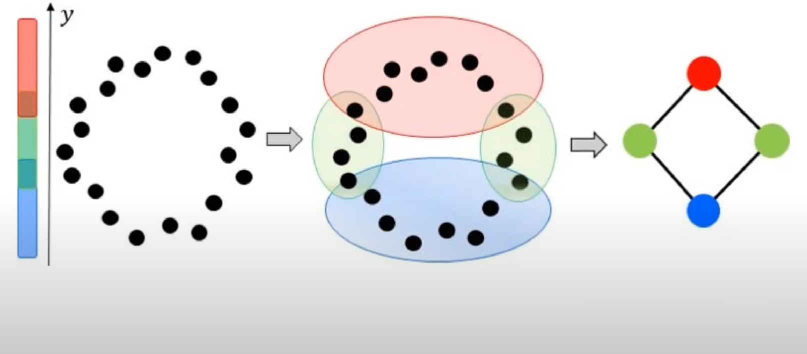
6th May 2022

Guide- Prof. Rekha Santhanam

Mapper (Singh, Memoli, Carlsson 2007)

Topological Methods for the Analysis of High Dimensional Datasets and 3d object recognition

Figure: Overview of the Mapper Technique [1]



[1] Taken from Dr. BalaKrishnamoorthy's slides

Mapper

Definition (Simplex)

A n-simplex $[v_0, v_1, .., v_n]$ is the smallest convex set in $\mathbb{R}^m$ containing $n+1$ points $v_0, v_1, .., v_n$ such that the vectors $v_1 - v_0$, $v_2 - v_0$,..,$v_n - v_0$ are linearly independent.
A 0-simplex is simply a point. A 1-simplex a line, a 2-simplex a solid triangle.

Mapper

Definition (Simplex)

A n-simplex $[v_0, v_1, .., v_n]$ is the smallest convex set in $\mathbb{R}^m$ containing $n+1$ points $v_0, v_1, .., v_n$ such that the vectors $v_1 - v_0$, $v_2 - v_0$,..,$v_n - v_0$ are linearly independent.
A 0-simplex is simply a point. A 1-simplex a line, a 2-simplex a solid triangle.

Definition (Face)

A m-face of a simplex $[v_0, v_1, .., v_n]$ is a subsimplex with $m+1$ vertices subset of the $v_i's$

Mapper

Definition (Simplex)

A n-simplex $[v_0, v_1, .., v_n]$ is the smallest convex set in $\mathbb{R}^m$ containing $n+1$ points $v_0, v_1, .., v_n$ such that the vectors $v_1 - v_0$, $v_2 - v_0$,..,$v_n - v_0$ are linearly independent.
A 0-simplex is simply a point. A 1-simplex a line, a 2-simplex a solid triangle.

Definition (Face)

A m-face of a simplex $[v_0, v_1, .., v_n]$ is a subsimplex with $m+1$ vertices subset of the $v_i's$

Definition (Simplicial Complex)

A simplicial complex is a collection of simplices such that if A is a simplex then each of its face is a simplex, and two $n+1$ simplices can share at most 1 n-face.

Mapper

Definition (Nerve)

Given a finite covering $\mathcal{U} = \{U_\alpha\}$ of a space X, define the nerve of the covering $\mathcal{U}$, $\mathcal{N}(\mathcal{U})$ to be the simplicial complex where a family $\{\alpha_0, \alpha_1, .., \alpha_k\}$ spans a k-simplex iff $U_{\alpha_0} \cap U_{\alpha_1} \cap .. \cap U_{\alpha_k} \neq \phi$

- Let $X = \mathbb{R}$ and $\mathcal{U} = \{(-\infty, -5), (-10, 10), (5, \infty)\}$. Then $\mathcal{N}(\mathcal{U}) = \{\{U_0, U_1\}, \{U_1, U_2\}\}$

Mapper

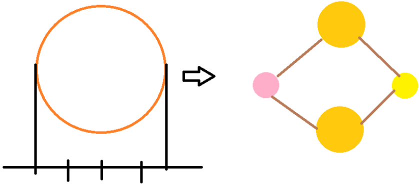
Definition (Mapper (Topological Version))

Given a space X and a continuous function $f : X \to Z$ with the space Z being equipped with a finite open cover $\mathcal{U} = \{U_\alpha\}$, $\{f^{-1}(U_\alpha)\}$ will be an open cover for X. Now, for each α, consider the decomposition of $\{f^{-1}(U_\alpha)\}$ into its connected components $\{f^{-1}(U_\alpha)\} = \cup_{i=1}^{j_\alpha} V(\alpha, i)$ (j_α is the total number of connected components of $\{f^{-1}(U_\alpha)\}$). Let $\mathcal{V} = \{V(\alpha, i)\}$. Then $\mathcal{N}(\mathcal{V})$ is the mapper output for the triplet $(X, f, \mathcal{U})$ or $\mathcal{M}(X, f, \mathcal{U}) = \mathcal{N}(\mathcal{V})$

Examples

Let $X = S^1$. Let $Z = \mathbb{R}$ and $\mathcal{U} = \{(-\infty, -0.5), (-1, 1), (0.5, \infty)\}$. We'll consider the mapper outputs for different $f : X \to Z$.

- let $f(\bar{x}) = \|\bar{x}\|$. Then $\mathcal{M}(X, f, \mathcal{U}) = \{V(2, 0)\}$
  Note that $f^{-1}(U_0) = f^{-1}(U_1) = \phi$ and $f^{-1}(U_2) = S^1$ which is connected as well.
- let $f((x, y)) = x$ Then $\mathcal{M}(X, f, \mathcal{U}) =$
  $\{\{V(0, 0), V(1, 0)\}, \{V(0, 0), V(1, 1)\}, \{V(1, 0), V(2, 0)\}, \{V(1, 1), V(2, 0)\}\}$



Mapper Implementation [Statistical Version]

Remark

To perform computations, the previous defn. is modified and we adopt clustering to partition a space into its connected components.

- Given a point cloud data X of N points, choose a filter function $f : X \to \mathbb{R}^k$ for some $k \in \mathbb{N}$.
- Now, partition the $Im(f)$, into overlapping bins. Over here we need to make two choices:
  i) resolution(r): Total no. of bins
  ii) gain(g): Related to % overlap between two bins, % overlap= 1-1/g.
- Suppose $U_1, U_2, .., U_r$ are the bins, then perform clustering on $f^{-1}(U_i) \forall i \in \{1, 2, .., r\}$. To perform clustering we need a distance matrix or some metric to compute the inter-point distances between points in the data.
- Add a $p-$simplex to the mapper output for every $(p+1)-$set intersection $f^{-1}(U_{\alpha_0}) \cap f^{-1}(U_{\alpha_1}) \cap ... \cap f^{-1}(U_{\alpha_p}) \neq \phi$

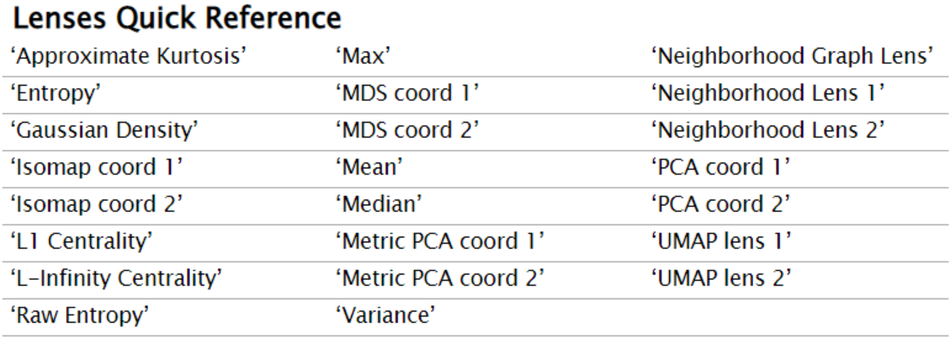
Filter Functions

The Mapper Output is highly dependent on the choice of filter functions. The filter functions can be chosen based on domain knowledge. Some possible filter functions are:

1. Projections: A dataset with n features can be projected down to a few of its columns.
2. PCA: Principal Component Analysis can be performed on its n features and the first few principal components can be considered.
3. Eccentricity: We first consider a distance matrix D of dim. $n \times n$ where d_{ij} is the distance between points i and j.Then for each point we compute the L^p norm.
   $E_p(x) = (\sum_{y \in X} d(x,y)^p)^{1/p}$. We have $k = 1$ in this case.
4. Density: Density estimation via the gaussian kernel $f_\epsilon(x) = C_\epsilon \sum_{y \in X} exp(-d(x,y)/\epsilon^2)$

Filter functions

Figure: Filter Functions provided by Ayasadi



Lenses Quick Reference

‘Approximate Kurtosis’	‘Max’	‘Neighborhood Graph Lens’
‘Entropy’	‘MDS coord 1’	‘Neighborhood Lens 1’
‘Gaussian Density’	‘MDS coord 2’	‘Neighborhood Lens 2’
‘Isomap coord 1’	‘Mean’	‘PCA coord 1’
‘Isomap coord 2’	‘Median’	‘PCA coord 2’
‘L1 Centrality’	‘Metric PCA coord 1’	‘UMAP lens 1’
‘L–Infinity Centrality’	‘Metric PCA coord 2’	‘UMAP lens 2’
‘Raw Entropy’	‘Variance’	

Application 1

i) Identification of Type 2 Diabetes subgroups through topological analysis of patient similarity [Li Li etal]

In this study, the authors developed a data-driven, topology-based approach to:

1) map the complexity of patient populations using clinical data from electronic medical records(EMRs) and
2) identify new, emergent Type 2 Diabetes patient subgroups with subtype-specific clinical and genetic characteristics.

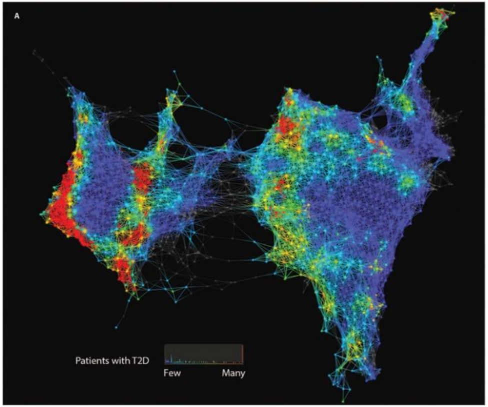
Mapper pipeline

i)Dataset: A dataset of 11,210 patients with several clinical variables on which analysis was performed. A separate Algorithm EMERGE was used to identify the T2D patients which resulted in 2,551 patients. A second analysis was conducted on these 2551 patients with 73 clinical variables.

ii)Filter functions: Two filter functions were used:
L-infinity Centrality- $L^{\infty}(x) = max_{y \in X} d(x, y)$ and the principal metric singular value decomposition

iii)Distance Metric: Cosine Similarity metric was used, $d(x, y) = \frac{x.y}{\|x\|\|y\|}$

Mapper Output



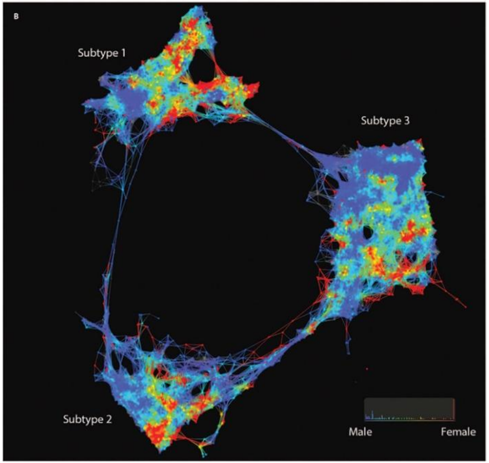


Fig1: Patient-Patient network with 11,210 patient record
Fig2: Patient-Patient network with 2551 T2D patients
Taken from Identification of type 2 diabetes subgroups through topological analysis of patient similarity

Results

- The Mapper output clearly gives us three distinct subgroups of T2d Patients with 762 (subtype 1), 617 (subtype 2), and 1096 (subtype 3) patients in each.

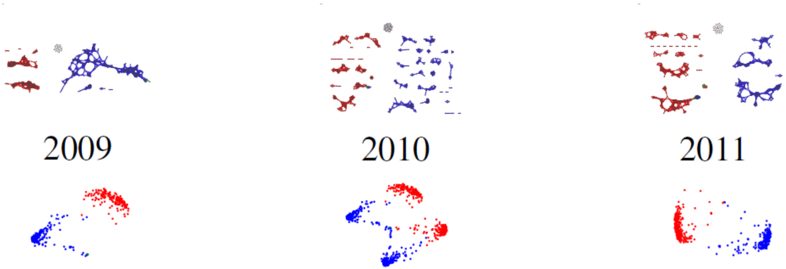
- With the help of the clusters clinical variables could be identified which were more specific to a particular subgroup. For example, 33 clinical variables were identified specific to subtype 1, 3 to subtype 2 and 11 to subtype 3.

- Through the study they were able to identify both the diseases that characterized the subtypes, as well as specific specific bio-markers. For example, subtype 1 was characterized by diabetic nephropathy and diabetic retinopathy; subtype 2 with cancer malignancy and cardiovascular diseases; and subtype 3 was with cardiovascular diseases, allergies and HIV infections.

Application 2

Topology of politics: voting connectivity in the US House of Representatives [Carlsson et al]
The aim of the paper is to highlight how topological methods can uncover hidden structures in data that classical machine learning methods fail to discover.

Figure: Mapper Output vs PCA output [1]



[1] Taken from The topology of politics [Carlsson et al]

Mapper Pipeline

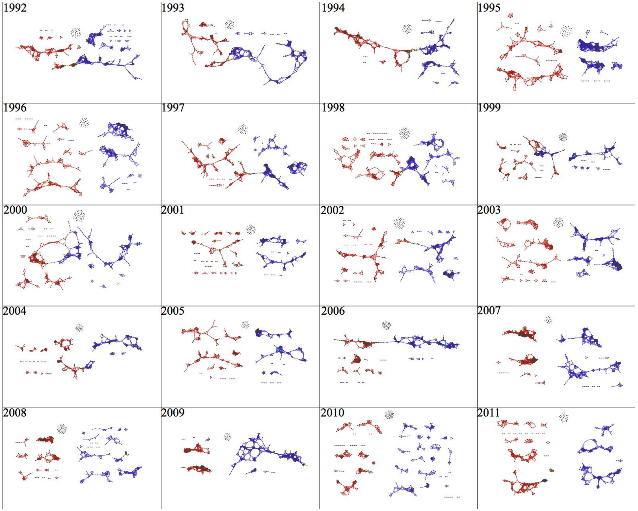
i)Dataset: Voting records from US HoR which take values Yes (1), No(-1) and others(0). Matrices are curated for each year with rows indicating representatives and columns representing the plenary votes. The number of rows stay between 435 and 447, whereas columns (except for a few years) stay between 444 and 691

ii)Filter functions: Principal metric singular value decomposition: the top two SVD co-ordinates computed using the correlation metric.

iii)Distance Metric: Correlation metric was used.

iV) resolution, r = 120 and gain = 4.5

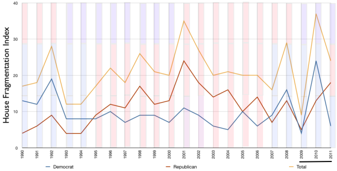
Mapper Output



Taken from The topology of politics [Carlsson et al]

Results

- One feature that's clear in the analysis is representatives voting along party lines.
- Authors point out to the finer structures that mapper exposes, the internal fragmentation and subgroups within the parties.
- Spikes in fragmentation count in the years 2001, 2008 and 2010. 2008 and 2010 had intense debates due to the economic slump and the Obama healthcare bill.



- Authors also talk about subgroups of the Republicans and Democrats who vote along similar lines (Central Group) is indicated by the mapper output.

Application 3

Exploring User Capability data with topological data analysis. [U Persad et al]

- The authors aimed to explore the global shape and sub-groupings (clusters of profiles) of people using data collected from the Cambridge Better Design Pilot Study.
- They wanted to better understand the data on human capability variation across populations which can support the approach of inclusive design.
- Mapper converts the high dimensional datasets into compact visual representations which the designers can understand and use.

Mapper Pipeline

i)Dataset: Data from 362 people regarding 39 variables consisting of sensory variables like near-vision, distance-vision and hearing at different volumes, motor variables like grip strength, getting out of a chair and cognitive variables like recall memory, numeracy.

ii)Filter functions: Neigbourhood Lens 1 and 2 (Ayasadi)

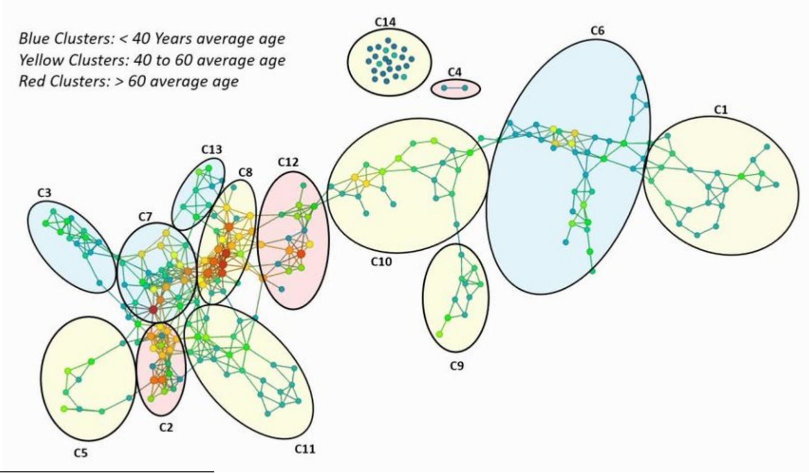
iii)Distance Metric: Norm Angle distance

$NormAngle(x, y) = \cos^{-1}(\frac{x^{trans}.y^{trans}}{\|x^{trans}\|\|y^{trans}\|})$ where $x^{trans} = (x - mean)/std.dev$

iV) resolution, r = 30 and gain = 2.5

Mapper Output

Clustered Mapper Output [1]


Blue Clusters: < 40 Years average age
Yellow Clusters: 40 to 60 average age
Red Clusters: > 60 average age
C14
C4
C6
C1
C13
C8
C12
C3
C7
C10
C9
C5
C2
C11

[1] Taken from Exploring User Capability Data with Topological Data Analysis [Persad et al]

Results

- Mapper was able to create a graph which when subjected to a graph clustering algorithm created 14 clusters, which give an indication of the structure in capability distribution across populations.
- The clusters provide information about the capability loss the group faces which can help designers design products that are more accessible, usable and easy to learn.
- The authors express that the pilot study can lead to larger scale data collection efforts as they have the means to identify the structures inherent in the data.

References

- Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition [Carlsson et al]
- Exploring User Capability Data with Topological Data Analysis[Persad et al]
- https://giotto-ai.github.io/gtda-docs/0.5.1/modules/mapper.html
- Identification of type 2 diabetes subgroups through topological analysis of patient similarity[Li Li et al]
- Extracting insights from the shape of complex data using topology [Lum et al]
- Topology of politics: voting connectivity in the US House of Representatives [Carlsson et al]