Clustering and mapper

Andrew J. Blumberg (blumberg@math.utexas.edu)

June 17th, 2014

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Overview

Goal of talk

Explain Mapper, which is the most widely used and most successful TDA technique. (At core of Ayasdi, TDA company founded by Gunnar Carlsson.)

Basic idea: perform clustering at different "scales", track how clusters change as scale varies.

Motivation:

1. Coarser than manifold learning, but still works in nonlinear situations.
2. Still retains meaningful geometric information about data set.
3. Efficiently computable (and so can apply to very large data sets).

Morse theory

Basic idea

Describe topology of a smooth manifold M using levelsets of a suitable function $h\colon M \to \mathbb{R}$.

- We recover M by looking at $h^{-1}((\infty, t])$, as t scans over the range of h.
- Topology of M changes at critical points of h.

Morse theory

Basic idea

Describe topology of a smooth manifold M using levelsets of a suitable function $h\colon M \to \mathbb{R}$.

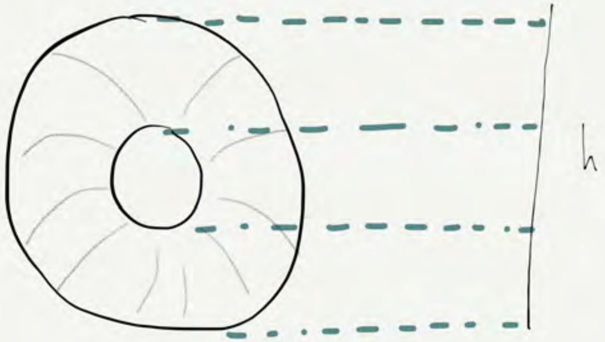
- We recover M by looking at $h^{-1}((\infty, t])$, as t scans over the range of h.
- Topology of M changes at critical points of h.

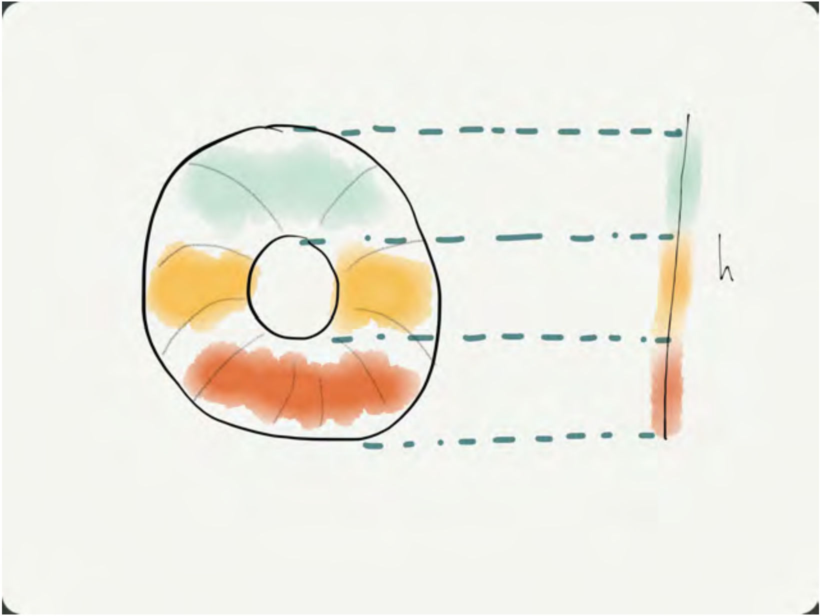
Morse theory

Basic idea

Describe topology of a smooth manifold M using levelsets of a suitable function $h\colon M \to \mathbb{R}$.

- We recover M by looking at $h^{-1}((\infty, t])$, as t scans over the range of h.
- Topology of M changes at critical points of h.


h


h

Reeb graphs

Convenient simplification:

1. For each $t \in \mathbb{R}$, contract each component of $f^{-1}(t)$ to a point.
2. Resulting structure is a graph.

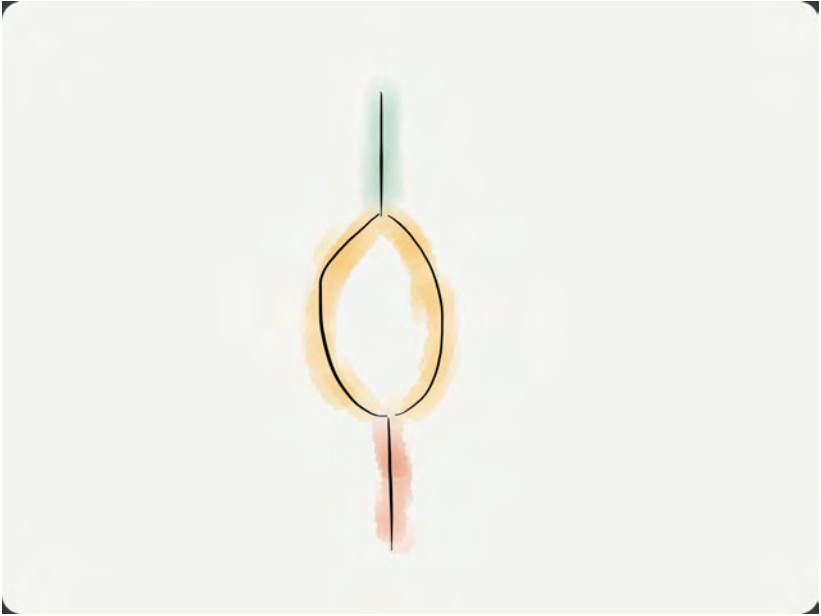
Convenient simplification:

1. For each $t \in \mathbb{R}$, contract each component of $f^{-1}(t)$ to a point.
2. Resulting structure is a graph.

Reeb graphs

Convenient simplification:

1. For each $t \in \mathbb{R}$, contract each component of $f^{-1}(t)$ to a point.
2. Resulting structure is a graph.



The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f: X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f\colon X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
   1. Clusters are vertices.
   2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f : X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f\colon X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f : X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f\colon X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f : X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f : X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Mapper

The mapper algorithm is a generalization of this procedure. [Singh-Memoli-Carlsson]

Assume given a data set X.

1. Choose a filter function $f \colon X \to \mathbb{R}$.
2. Choose a cover U_α of X.
3. Cluster each inverse image $f^{-1}(U_\alpha)$.
4. Form a graph where:
    1. Clusters are vertices.
    2. An edge connects two clusters C and C' if both $U_\alpha \cap U_{\alpha'} \neq \emptyset$ and $C \cap C' \neq \emptyset$.
5. Color vertices according to average value of f in the cluster.

Filter functions

Clearly, choice of filter function is essential.

- Some kind of density measure.
- A score measure difference (distance) from some baseline.
- An eccentricity measure.

Filter functions

Clearly, choice of filter function is essential.

- Some kind of density measure.
- A score measure difference (distance) from some baseline.
- An eccentricity measure.

Filter functions

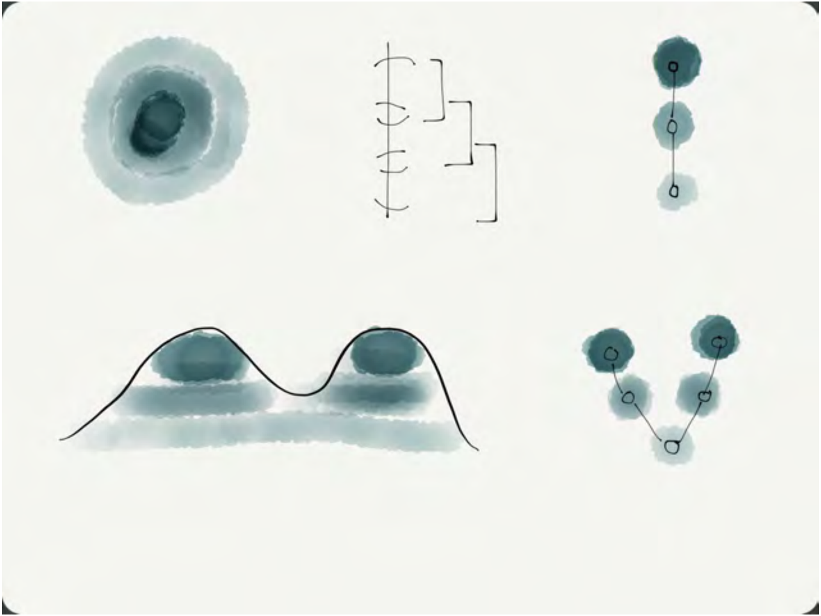
Clearly, choice of filter function is essential.

- Some kind of density measure.
- A score measure difference (distance) from some baseline.
- An eccentricity measure.

Filter functions

Clearly, choice of filter function is essential.

- Some kind of density measure.
- A score measure difference (distance) from some baseline.
- An eccentricity measure.



Breast cancer example

Highly successful example of real data analysis. [Nicolau, Carlsson, Levine]

- Working with vectors of gene expression data.
- Distance metric is correlation.
- Filter is a measure of (unsigned) deviation of expression from normal tissue.

Results identify previously unknown c-MYB+ region, which are very different from normal tissue but have very high survival rates.

Breast cancer example

Highly successful example of real data analysis. [Nicolau, Carlsson, Levine]

- Working with vectors of gene expression data.
- Distance metric is correlation.
- Filter is a measure of (unsigned) deviation of expression from normal tissue.

Results identify previously unknown c-MYB+ region, which are very different from normal tissue but have very high survival rates.

Breast cancer example

Highly successful example of real data analysis. [Nicolau, Carlsson, Levine]

- Working with vectors of gene expression data.
- Distance metric is correlation.
- Filter is a measure of (unsigned) deviation of expression from normal tissue.

Results identify previously unknown c-MYB+ region, which are very different from normal tissue but have very high survival rates.

Breast cancer example

Highly successful example of real data analysis. [Nicolau, Carlsson, Levine]

- Working with vectors of gene expression data.
- Distance metric is correlation.
- Filter is a measure of (unsigned) deviation of expression from normal tissue.

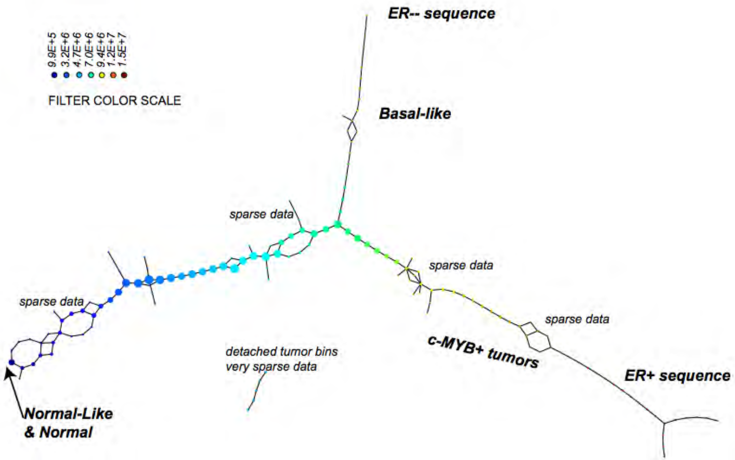
Results identify previously unknown c-MYB+ region, which are very different from normal tissue but have very high survival rates.

Breast cancer example

Highly successful example of real data analysis. [Nicolau, Carlsson, Levine]

- Working with vectors of gene expression data.
- Distance metric is correlation.
- Filter is a measure of (unsigned) deviation of expression from normal tissue.

Results identify previously unknown c-MYB+ region, which are very different from normal tissue but have very high survival rates.



9.9E+5
3.2E+6
4.7E+6
7.0E+6
9.4E+6
1.2E+7
1.5E+7
FILTER COLOR SCALE
ER-- sequence
Basal-like
sparse data
sparse data
sparse data
sparse data
c-MYB+ tumors
ER+ sequence
detached tumor bins
very sparse data
Normal-Like
& Normal

NBA example

Clever example of application to sports analytics. [Alagappan]

- Data set consists of vectors of statistics (points scored, rebounds, etc.).
- Distance metric is Euclidean.
- Filter is points per minute.

Results identify many “new” positions.

NBA example

Clever example of application to sports analytics. [Alagappan]

- Data set consists of vectors of statistics (points scored, rebounds, etc.).
- Distance metric is Euclidean.
- Filter is points per minute.

Results identify many “new” positions.

NBA example

Clever example of application to sports analytics. [Alagappan]

- Data set consists of vectors of statistics (points scored, rebounds, etc.).
- Distance metric is Euclidean.
- Filter is points per minute.

Results identify many "new" positions.

NBA example

Clever example of application to sports analytics. [Alagappan]

- Data set consists of vectors of statistics (points scored, rebounds, etc.).
- Distance metric is Euclidean.
- Filter is points per minute.

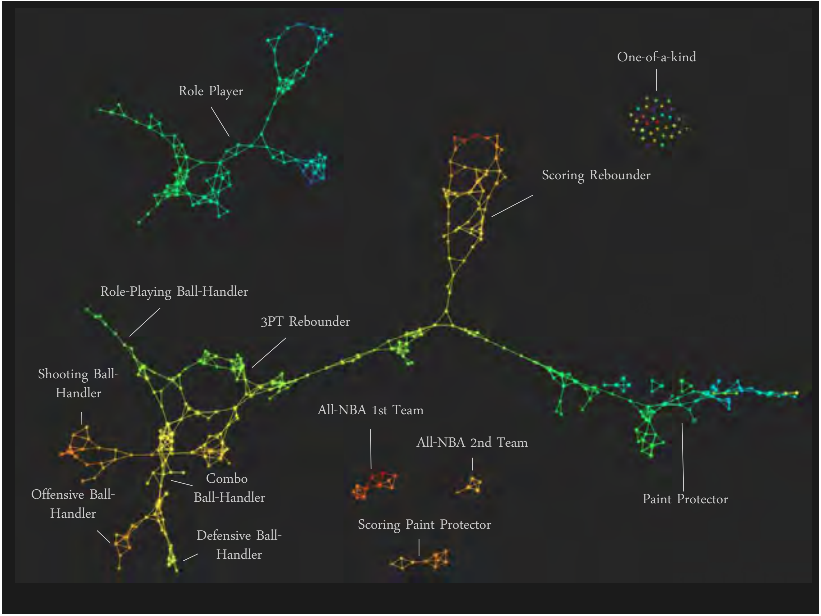
Results identify many "new" positions.

NBA example

Clever example of application to sports analytics. [Alagappan]

- Data set consists of vectors of statistics (points scored, rebounds, etc.).
- Distance metric is Euclidean.
- Filter is points per minute.

Results identify many "new" positions.


One-of-a-kind
Role Player
Scoring Rebounder
Role-Playing Ball-Handler
3PT Rebounder
Shooting Ball-Handler
All-NBA 1st Team
All-NBA 2nd Team
Combo Ball-Handler
Offensive Ball-Handler
Paint Protector
Scoring Paint Protector
Defensive Ball-Handler

Summary

Claim

Mapper can be successfully applied to analysis of geometric structures in large data sets from a wide variety of domains.

- Key idea: clustering across "scales", represent relationships between clusters as scale varies
- Choice of filter function(s) is critical to successful aplication.

Summary

Claim

Mapper can be successfully applied to analysis of geometric structures in large data sets from a wide variety of domains.

- Key idea: clustering across “scales”, represent relationships between clusters as scale varies
- Choice of filter function(s) is critical to successful aplication.

Summary

Claim

Mapper can be successfully applied to analysis of geometric structures in large data sets from a wide variety of domains.

- Key idea: clustering across "scales", represent relationships between clusters as scale varies
- Choice of filter function(s) is critical to successful aplication.