Департамент образования и науки Брянской области

Государственное бюджетное профессиональное образовательное учреждение

“Новозыбковский профессионально-педагогический колледж”

ИНДИВИДУАЛЬНЫЙ ПРОЕКТ

Изучение инструментов топологического анализа данных и его применения в современном data science

Сычев Арсений Евгеньевич, Л-22

Научный руководитель:

Саросек Сергей Михайлович

Новозыбков, 2021

Оглавление

Введение

Каждый день человечество генерирует примерно 2,5 квинтиллиона байт различных данных. Они создаются буквально при каждом клике и пролистывании страницы, не говоря уже о просмотре видео и фотографий в онлайн-сервисах и соцсетях.

Наука о данных появилась задолго до того, как их объемы превысили все мыслимые прогнозы. Отсчет принято вести с 1966 года, когда в мире появился Комитет по данным для науки и техники — CODATA. Его создали в рамках Международного совета по науке, который ставил своей целью сбор, оценку, хранение и поиск важнейших данных для решения научных и технических задач. В составе комитета работают ученые, профессора крупных университетов и представители академий наук из нескольких стран, включая Россию.

Сам термин Data Science вошел в обиход в середине 1970-х с подачи датского ученого-информатика Петера Наура. Согласно его определению, эта дисциплина изучает жизненный цикл цифровых данных от появления до использования в других областях знаний. Однако со временем это определение стало более широким и гибким. Data Science имеет множество областей и ответвлений таких как машинное обучение, глубокое обучение, большие данные, искусственный интеллект и топологический анализ данных. Среди перечисленных, Топологический Анализ Данных представляет для меня наибольший интерес, так как в данной области необходимы знания алгебраической топологии, статистики и других областей математики. Топологический Анализ Данных использует одни из последних достижений математики для нахождения глубоких и неочевидных закономерностей.

Объект данного исследования - процесс изучения методов исследования данных с помощью топологического анализа данных.

Предмет исследования - применение алгоритма Mapper, топологических инвариантов для исследования набора данных.

Цель исследования – достижение понимания выявления глубоких закономерностей в наборах данных, основанных на топологических инвариантах пространств.

1. История развития и применения топологического анализа данных.

Топологический анализ данных – это достаточно новая область анализа данных.

Основной метод топологического анализа данных - замена набора элементов данных некоторым семейством симплициальных комплексов в соответствии с параметром близости. Анализ этих топологических комплексов с помощью алгебраической топологии, а конкретно новой теорией персистентных гомологий.

Перекодировка устойчивой гомологии набора данных в параметризованную версию чисел Бетти, называемую баркодом.

Одним из основных понятий в TDA (Топологический анализ данных) является облако точек - набор вершин в трёхмерной системе координат Эти вершины, как правило, определяются координатами X, Y и Z и, как правило, предназначены для представления внешней поверхности объекта.

Данные часто представлены множеством точек в Евклидовом пространстве E_n, форма которого отражает описываемый данными феномен.

Реальные трехмерные объекты могут представляться в виде облака точек. Например лазером отмечаются отдельные точки и их неструктурированный набор служит представлением объекта в компьютере. Облаком точек считается любой (возможно зашумленный) набор точек в E_n или проекций точек в более низкой размерности.

В компьютерной графике и статистике есть различные методы построения прообразов по проекциям. Топологический анализ данных предназначен для пространств высоких размерностей или слишком искривленных чтобы создавать по ним плоские проекции.

Для преобразования облака точек в метрическом пространстве в целостный объект точки используются в качестве вершин графа ребрам которого приписаны расстояния, затем граф превращается в симплициальный комплекс и изучается средствами алгебраической топологии.

Топологический анализ данных активно используется в современном data science. Ниже приведены примеры исследований в области медицины, в ходе которых специалисты выявили новые взаимосвязи организма, что позволит эффективнее распознавать и лечить многие заболевания

В 70-х годах прошлого века в Стэнфордском Университете было проведено исследование. Были исследованы 145 пациентов-диабетиков.

У каждого из них было измерено 6 медицинских параметров(возраст, относительный вес, уровень глюкозы в плазме натощак, площадь под кривой глюкозы в плазме для трехчасового теста на толерантность к глюкозе (OGTT), площадь под кривой плазменного инсулина до этого теста, базовый плазменный уровень глюкозы).

Данные полученные в ходе исследования представляли собой набор точек пространства R^6.

Позже Г.М.Ривен и Р.Г.Миллер применили метод поиска наилучшей проекции к этим данным и нашли "наилучшую проекцию" этого множества на трехмерное пространство (см. рис. 1).

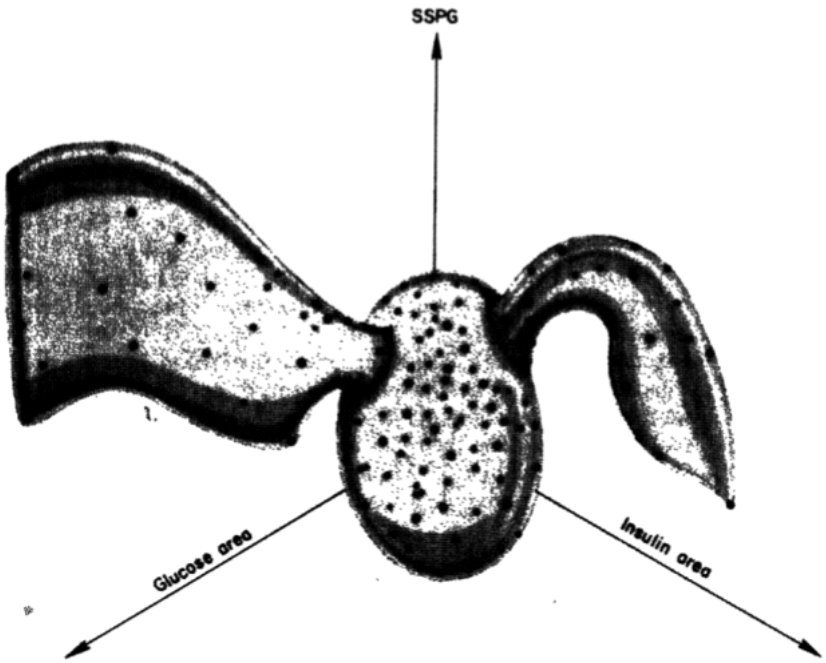


Рисунок 1. Пример проекции облака точек пространства R^6 в пространство R^3

Миллер и Ривен заметили, что в этом множестве данных можно выделить центральное ядро и два отростка, из него выходящих.

Пациенты в каждом из отростков оказались пациентами страдающими от диабетов первого и второго типа. Данное открытие позволило ученым лучше понять разница между данными заболеваниями.

1)Распознавание опухолей с помощью чисел Бетти:

В этой статье используются числа Бетти для автоматического распознавания рака прямой кишки. Раньше это делалось врачом “на глаз” путем анализа большого количества микро-фотографий прямой кишки. Для автоматизации было использовано машинное обучение. Данный подход дал точность 91%. Совмещение топологического анализа данных и машинного обучения дало точность 94%. показано сравнение двух образцов. Слева – с опухолью, справа –без.(см. рис. 2)

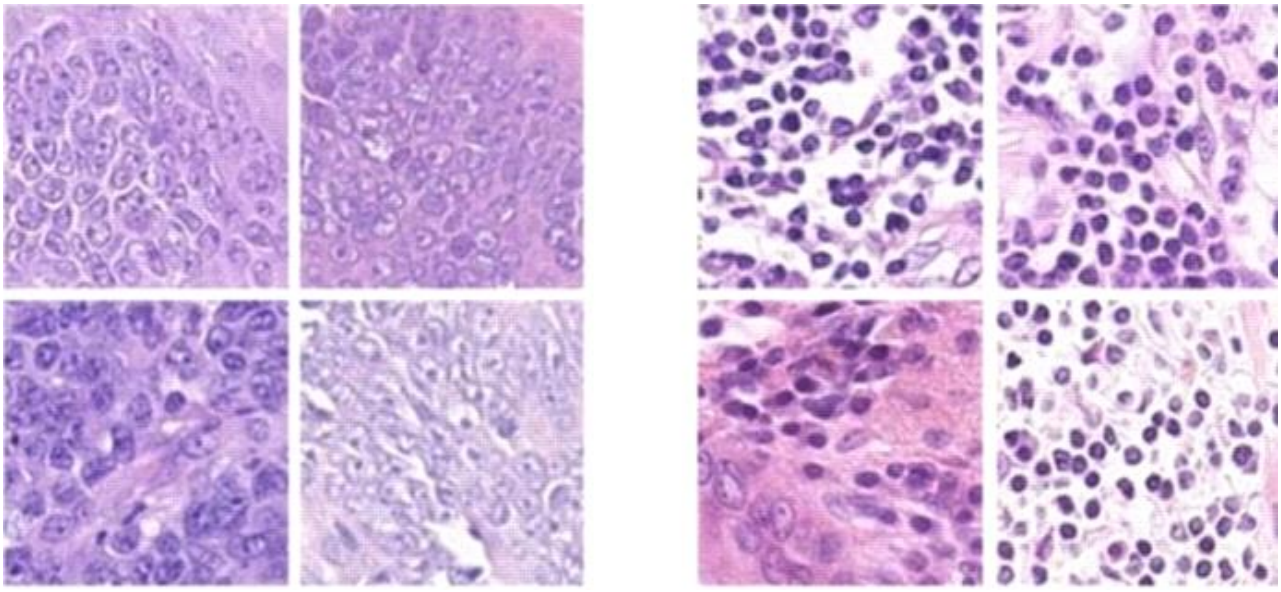


Рисунок 2. сравнение двух образцов. Слева – с опухолью, справа –без.

Алгоритм:

-Яркость пикселя в rgb-формате – (r+g+b)/3.

-Для картинки рассматриваем множество пикселей B_a, которые не ярче 0 $\leqslant a \leqslant 255$:

$$B_0 \subseteq B_1 \subseteq B_2 \subseteq \ldots \subseteq B_{255}$$

Рассматриваем соответствующие числа Бетти $\beta_0(B_a)$ и $\beta_1(B_a)$

(см. рис.3)

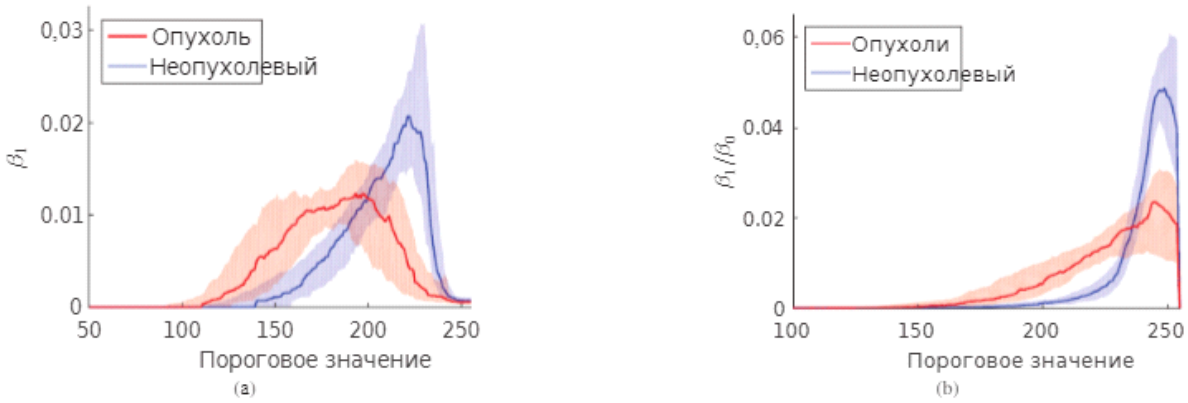


Рисунок 3. Распределение времен жизни(слева) и рождений(справа)

Функциональные сети головного мозга и псилоцибин

Идея данного исследования состоит в том, что мозг разбивается на 194 участка. При помощи МРТ измеряется активность выбранных участков и вычисляются корреляции активностей.

По этим данным строется взвешенный граф, вершины которого - участки мозга, ребра существуют между коррелирующими участками, вес ребра определяется при помощи корреляции.

Далее строится соответствующий фильтрованный комплекс и вычисляются его первые персистентные гомологии.

Фильтрация комплекса происходит по значению его ребер(т.е. ребра с малой корреляцией

Обычно в качестве веса w(X, Y) берут значение 1 - r(X, Y)^2

Далее строится персистентная диаграмма, которая дает информацию о связях между активностью разных участков

Через π обозначается время жизни персистентной гомологии. через β - время рождения.(см рис.4)

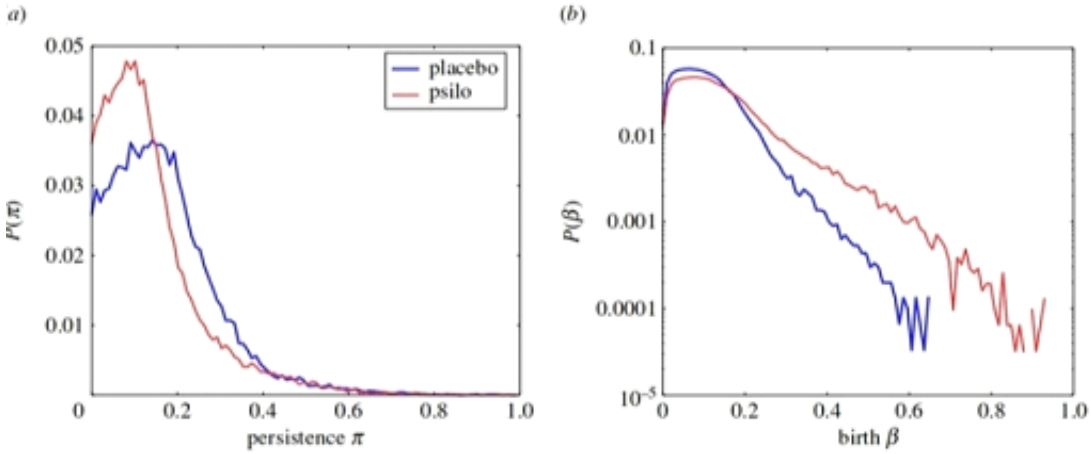


рисунок 4. распределение времен жизни(слева) и рождений (справа)

Исходя из графиков заметно, что под действием псилоцибина скорость появление новых циклов падает медленнее, но циклы при этом являются менее устойчивыми.

Авторы рассматривают базис персистентных гомологий $\{g_i\}$

и определяют новые веса ребер:

$$w^{\pi}(e) = \sum_{g_i | e \in g_i} \pi(g_i).$$

То есть, ребро имеет большой вес, если суммарное время жизни циклов, проходящих через него большое.

Получившийся взвешенный граф называется персистентным каркасом.

На рисунке 5 показаны персистентные каркасы, с выкинутыми ребрами веса меньше 80

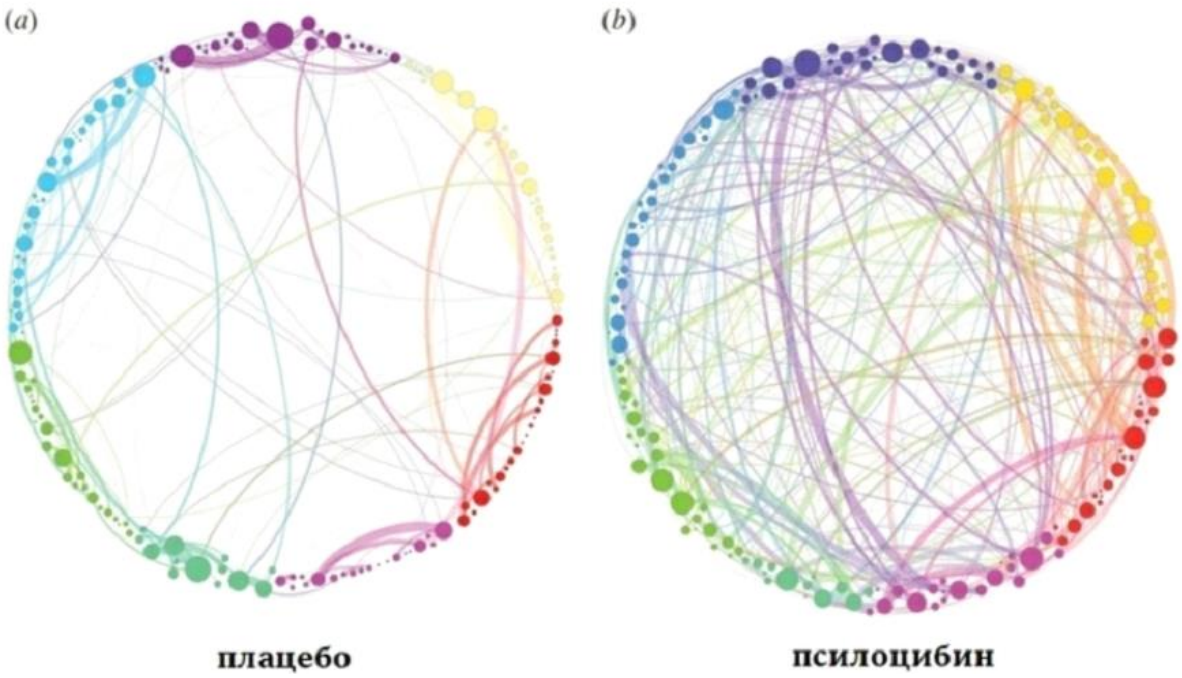
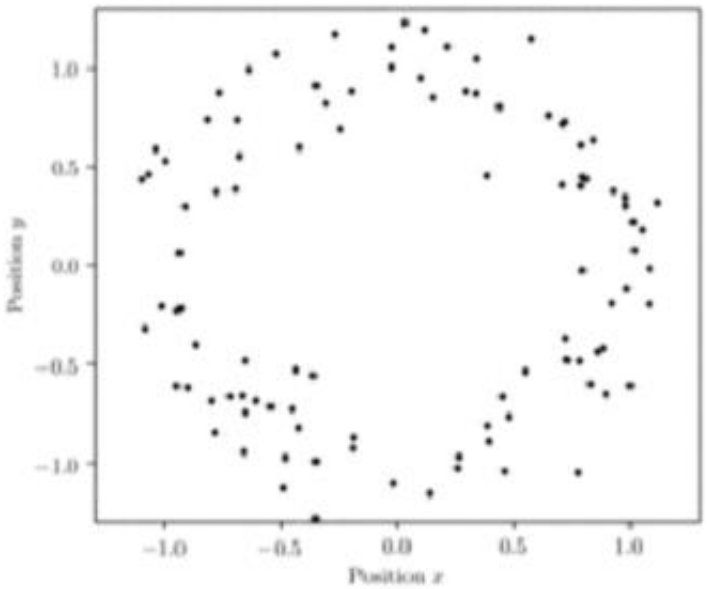


Рисунок 5. Персистентные каркасы с выкинутыми ребрами меньше 80

2. Основные понятия топологического анализа данных

Рассмотрим облако точек.



Видно, что это облако по форме близко к окружности. Если размерность была бы выше , мы не смогли бы "на глаз" оценить форму.

Топологический анализ данных - наука, которая изучает "форму"

многомерных облаков точек методами, устойчивыми к шуму. В том числе при помощи методов алгебраической топологии

На метрических пространствах есть отношение изометричности. Оно сохраняет все расстояния.

Если необходимо работать с многомерными пространствами, так, чтобы созданная модель была устойчива к шуму(отклонениям), нужно рассматривать более слабые отношения эквивалентности.

Примером такого отношения эквивалентности является гомеоморфность.

Еще более слабым отношением эквивалентности является гомотопическая эквивалентность.

Метрическое пространство - пара (M, d), где M - множество и d: M x M -> $R_{>0}$, которая называется расстоянием или метрикой, такая что:

d(x,y) = 0 <=> x=y

d(x, y) = d(y, x)

d(x, y) + d(y, z) ≧ d(x, z)

пример метрического пространства:

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

рисунок 6. R^n с евклидовым расстоянием, которое вычисляется по данной формуле

Если X принадлежит M = (M, d) - метрическое пространство, то метрику можно сузить на X и получить метрику d: X x X -> $R_{>0}$ на X.

Любое подмножество X, принадлежащее R^n естественным образом является метрическим пространством.

Если X - любое множество, то на нем можно задать дискретную метрику:

$$d(x, y) = \begin{cases} 1, & \text{если } x \neq y \\ 0, & \text{если } x = y \end{cases}$$

C([a, b]) - множество непрерывных функций f: [a, b] -> R.

На C([a, b]) есть равнометрная метрика.

$d(f, g) = \max\{|f(x) - g(x)|: x \in [a, b]\}$.

на рис.7 представлено графическое представление такой метрики.

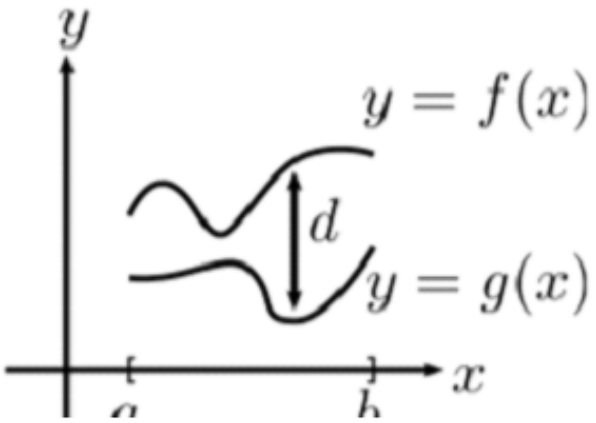


рисунок 7. графическое представление метрики на C([a,b])

Отображение f: M -> N называется изометрией, если для любых x y ∈ M

$d_M(x, y) = d_N(f(x), f(y))$.

Изометрия инъективна.

Глобальная изометрия - биективная изометрия

Два метрических пространства называются изометричными, если между ними есть глобальная изометрия.

Изометричность - максимально "сильное" отношение эквивалентности на метрических пространствах.

Открытый шар в M с центром в p ∈ M и радиусом r > 0 определяется как

$B_r(p) = \{x \in M | d(p,x) < r\}$

Если M принадлежит R^n, то открытый шар в M - это пересечение обычного открытого шара в R^n с M.

Если X - множество, то через P(X) обозначается множество его подмножеств.

Топологическое пространство - пара (X, O), где O принадлежит P(X), такие, что

$X \in O$ и пустое множество принадлежит O.

$U, V \in O \Rightarrow U \cap V \in O$.

O - называется топологией на X.

Напрерывное отображение f: X-> Y называется гомеоморфизмом, если оно биективно и обратное f^{-1}: Y ->X непрерывно.

Гомеоморфизм задает биекцию между Ox и O_y

На рисунке 8 представлено графическое представление изоморфизма.

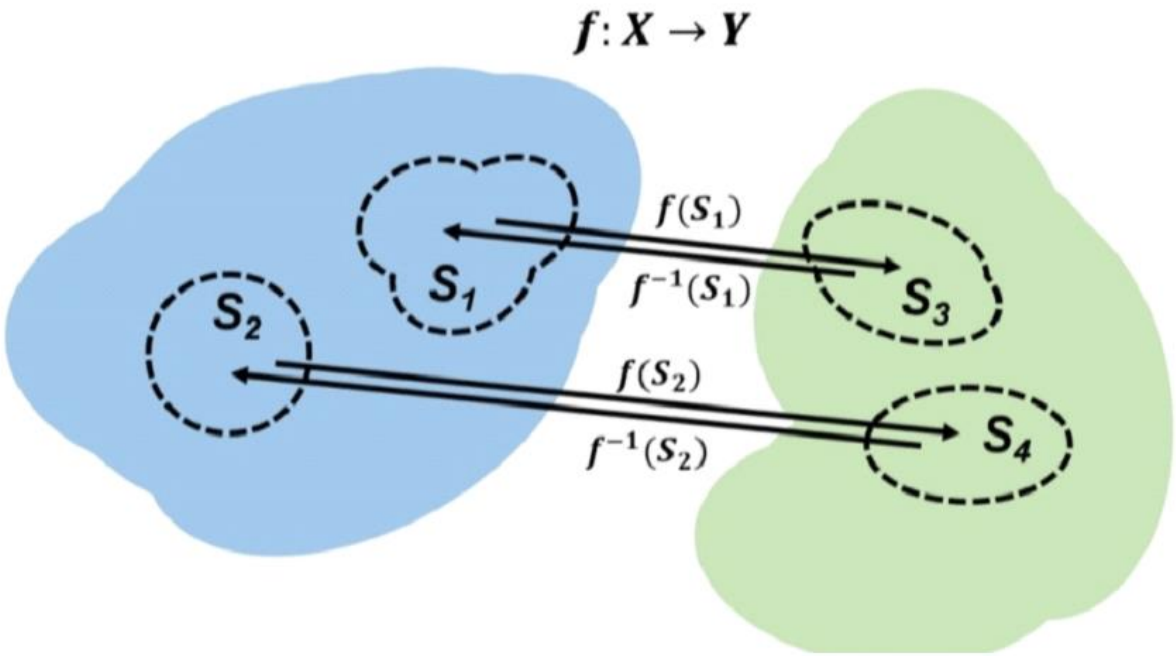


рисунок 8. графическое представление изоморфизма.

Пусть T - топологическое пространство.

Обозначим через I отрезок [0,1].

непрерывное отображение G:I -> T называется путем в T.

G(0) - начало пути, G(1) - конец пути.

Две точки T можно соединить путем, если есть путь с началом в одной точке и концом в другой.

Соответствующие классы эквивалентности называются компонентами линейной связности.

Компонента точки x обозначается через [x].

Множество компонент T оозначается через $\pi_0(T)$

если компонента одна, то пространство линейно-связное.

Если X, Y - топологические пространства, то их декартово произведение X x Y снабжается топологией произведение $O_{X \times Y}$, в которой открытыми

множествами являются обьединения любых семейств множеств вида U x V, где U открыто в X и V открыто в Y.

Пусть f, g: X->Y - два непрерывных отображения.

Грубо говоря гомотопия - непрерывное семейство непрерывных отображений f_t : X -> Y, где t принадлежит [0;1], которое соединяет

$f = f_0$ и $g = f_1$.

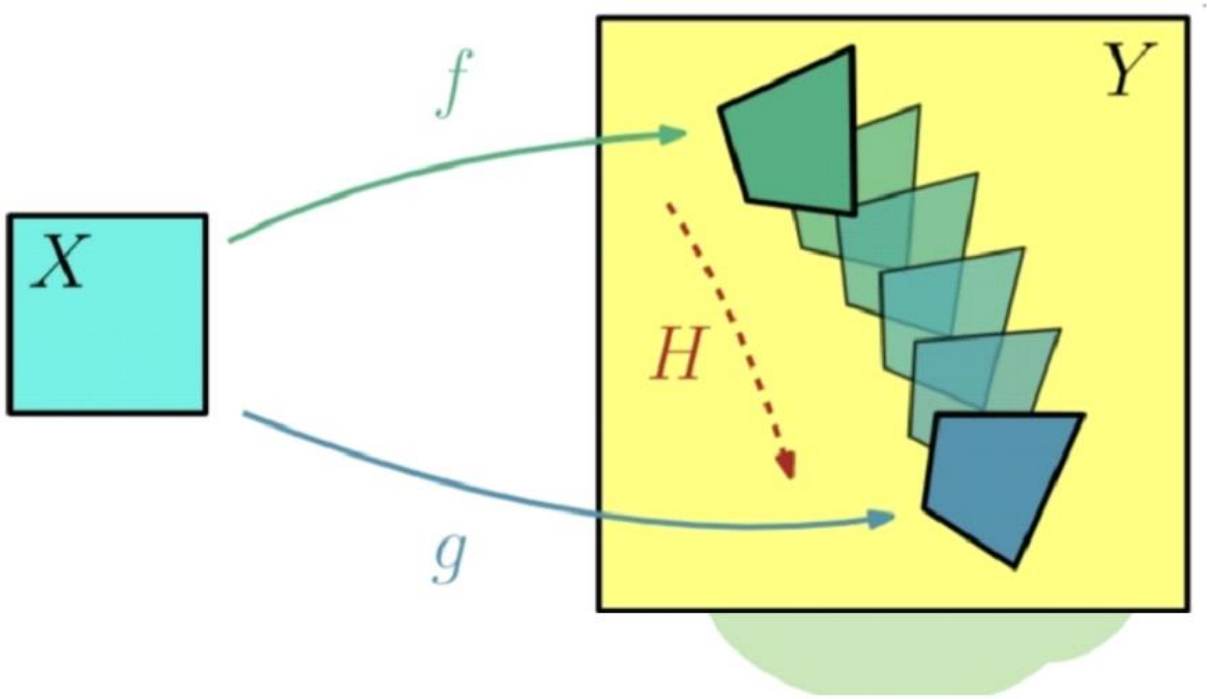


Рисунок 9. Визуализация гомотопии(1).

Гомотопия из f в g - это непрерывное отображение

H: X x I -> Y

такое, что f(x) = H(x, 0) и g(x) = H(x, 1).

Тогда можно обозначить $f_t(x) = H(x, t)$.

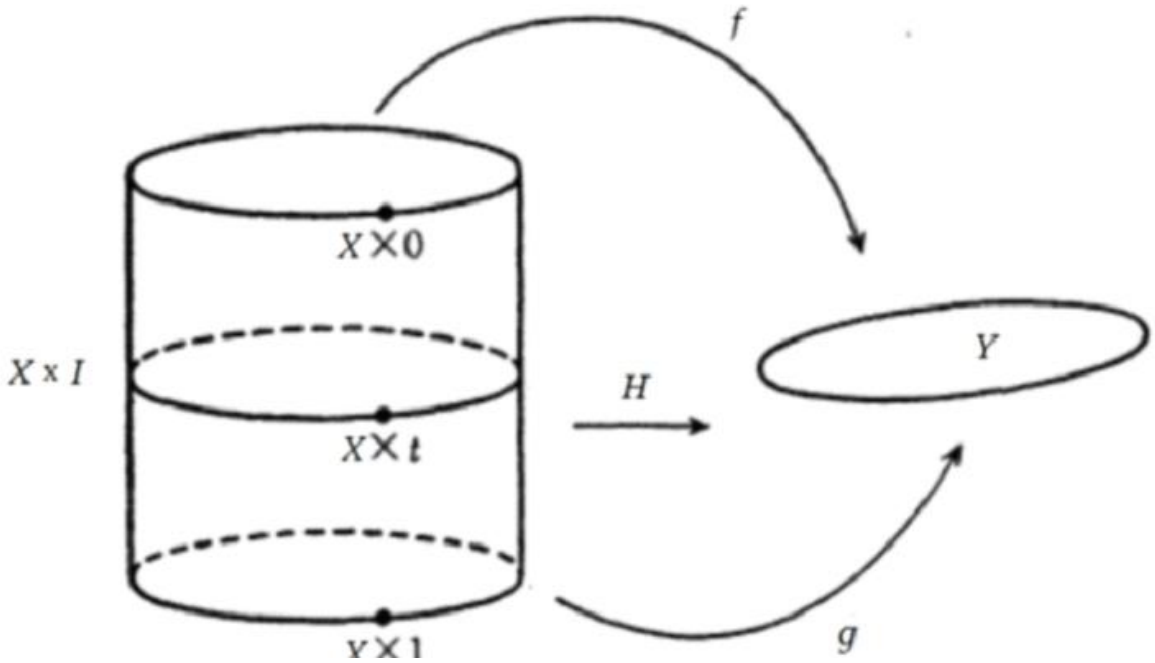


Рисунок 10. Визуализация гомотопии(2).

f и g называют гомотопными: f ~ g.

Отношение гомотопности - это отношение эквивалентности.

Множество классов эквивалентности непрерывных отображений обозначается [X, Y].

Грубо говоря, гомотопия - непрерывная деформация одного отображения на другое.

$[S^1, S^1]$ - бесконечно. и между $[S^1, S^1]$ и Z есть естественная биекция.

Два пространства X, Y называются гомотопически эквивалентными, если существуют непрерывные отображения

f: X -> Y, g: Y -> X

такие, что:

$$fg \sim gf \sim id$$

пространство стягиваемое, если оно гомотопически эквивалентно точке. Люое выпуклое подмножество R^n стягиваемо. Если пространство X стягиваемое, то [X, X] одноэлементно.

S^1 не стягиваемое

S^1 гомотопически эквивалентно $R^2 \setminus \{0\}$.

Преимуществом гомотопической эквиваленции является ее большая устойчивость к шуму.

Простейший инвариант гомотопической эквивалентности - гомологии.

n-симплекс - n-мерное обобщение треугольника и тетраэдра.

n-симплекс в R^d - выпуклая оболочки n+1 - ой точки в общем положении(не лежат в одном n-1-мерном аффинном подпространстве).

0-симплекс - точка.

1-симплекс - отрезок.

2-симплекс - треугольник.

3-симплекс - тетраэдр.

Грань n-симплекса - это k-симплекс, который является выпуклой оболочкой некоторого подмножества вершин этого симплекса.

Топологический симплициальный комплекс K - это множество симплексов в R^d такое, что

Для каждого симплекса из K его грани тоже лежат в K.

Пересение любых двух симплексов a, b ∈ K, либо пусто, либо является гранью и a и b.

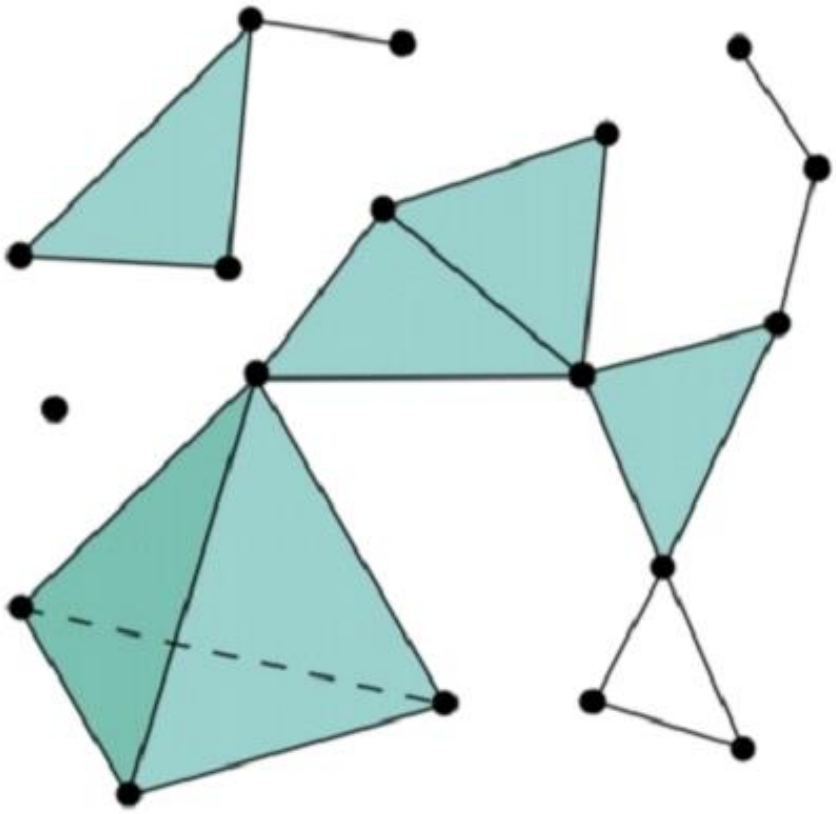


Рисунок 11. Пример симплекса

По топологическому симплициальному комплексу можно построить топологическое пространство - объединение его симплексов. Такие пространства тоже называют симплициальными комплексами.

Многие интересные пространства гомеоморфны симплициальным комплексам.

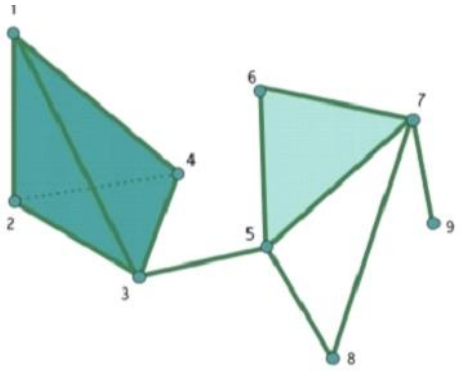
Например, сфера - это граница тетраэдра или икосаэдра.

Все пространства, которые могут пригодиться в топологическом анализе данных, гомотопичеки эквивалентных симплициальным комплексам.

Абстрактные симплициальные комплексы - это комбинаторные аналоги топологических симплициальных комплексов.

Абстрактный симплициальный комплекс - это множество конечных непустых множеств K такое, что если X ∈ K и множество Y лежащее в X и не равная пустому множеству, то Y принадлежит K.

Пример:



$$K = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\}$$
$$\{1,2\},\{1,4\},\{1,3\},\{2,3\},\{2,4\},\{3,4\},\{3,5\}$$
$$\{5,6\},\{5,7\},\{5,8\},\{6,7\},\{7,8\},\{7,9\},$$
$$\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\},\{5,6,7\},$$
$$\{1,2,3,4\}\}$$

n+1-элементы симплекса K называются n-симплексами.

Пусть F- поле.

V, U - векторные пространства над F.

Отображение f: V -> U называется линейным, если

$f(av) = af(v)$ и $f(v_1 + v_2) = f(v_1) + f(v_2)$.

$Im(f) = \{u \in U \mid \exists\ v \in V\ f(v) = u\}$

$ker(f) = \{v \in V \mid f(v) = 0\}$

$dim(Im(f)) + dim(ker(f)) = dim(V)$

Если U - подпространство векторного пространства V, то через V/U обозначается фактор-пространство.

$$V/U = \{v + U | v \in V\}$$

Есть сюрьективное линейное отображение f: V -> V/U,

которое задается формулой f(v) = v + U, которое называется канонической проекцией.

$$\ker(f:V \to V/U) = U$$

Если $r < s$, и $X_s \subseteq X_r$, то существует линейная отображения $H_n(X_r) \to H_n(X_s)$.

Если рассматривать не только числа Бетти, но и сами группы гомологий вместе с линейными отображениями, то можно извлечь гораздо больше информации.

n-ые персистентные гомологии облака точек X - это совокупность векторных пространств

$H_n(X_r)$ по всем числам $r > 0$ линейных отображений

$H_n(X_r) \to H_n(X_s)$ по всем парам чисел $0<r<s$

Клика графа - любое подмножество его вершин в котором все пары различных вершин соединены.

Одноэлементное множество считается кликой.

Кликовый комплекс X(G) графа G - это абстрактный симплиц. комплекс, симплексы которого - клики. Взвешенный граф - пара $(G, w:G_1 \to R)$.

G_a - граф G, из которого выкинули ребра веса больше a.

$(X(G_a)), a \in R$ - фильтрованный комплекс.

$H_n(X(G_a))$ - персистентные гомологии взвешенного графа (G, w)

Цепной комплекс $C_.$ – последовательность векторных пространств C_n, $n \in R$, и линейных отображений d_n: C_n -> C_{n-1}(называемых дифференциалами).

$$\ldots \to C_2 \to C_1 \to C_0 \to C_{-1} \to \ldots$$

Композиция последовательных дифференциалов равна нулю, т.е.

$$d_n d_{n+1} = 0$$

$Z_n(C_.) = \ker(d_n)$ – n-ая группа циклов $C_.$

$B_n(C_.) = \mathrm{Im}(d_{n+1})$ – n-ая группа границ $C_.$

$B_n(C_.)$ принадлежит $Z_n(C_.)$ и $Z_n(C_.)$ принадлежит C_n.

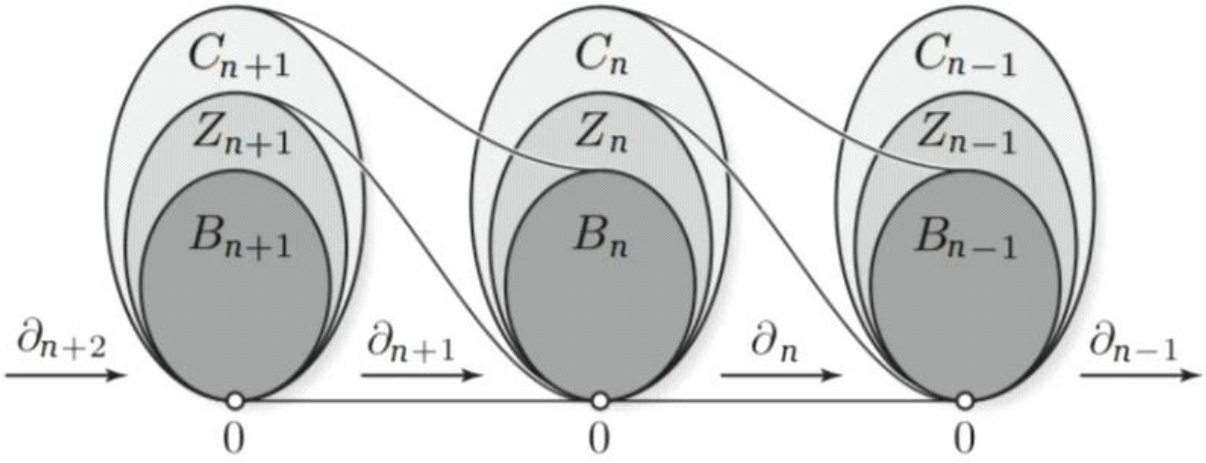


Рисунок 12. Визуализация строения комплекса.

n-ая группа гомологий цепного комплекса определяется как фактор-группа пространства циклов по пространству границ.

$$H_n(C_.) = Z_n(C_.)/B_n(C_.)$$

H_n измеряет то, насколько Z_n отличается от B_n.

Пусть есть два цепных комплекса C. И D. .Морфизм цепных комплексов(цепное отображение) f. : $C_\cdot$ -> D. – это последовательность линейных отображений f_n:C_n -> D_n таких, что $f_{n-1}d_n^C = d_n^D f_n$

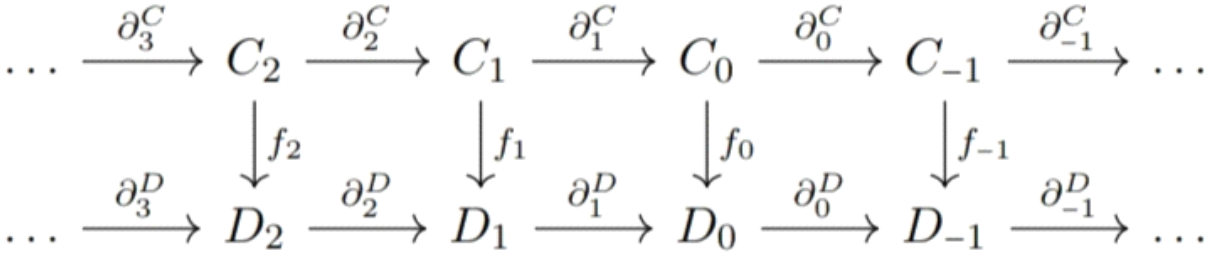


Рисунок 13.Визуализация цепного отображения

Изоморфизм цепных комплексов – морфизм, каждая компонента которого - изоморфизм

Если C. – цепной комплекс, у которого конечное число нулевых компонент и все они имеют конечную размерность, то справедлива формула, представленная на рисунке 14.

$$\sum_i (-1)^i \dim(C_i) = \sum_i (-1)^i \dim(H_i(C_\bullet)).$$

Рисунок 14. Эйлерова характеристика X(C.) комплекса C.

график Риба - это математический объект, отражающий эволюцию наборов уровней вещественнозначной функции на многообразии.

3.Описание набора данных для построение симплициального комплекса.

Характеристики вычисляются на основе оцифрованного изображения тонкоигольного аспирата (FNA) новообразования груди. Они описывают характеристики ядер клеток, представленных на изображении.

В трехмерном пространстве описано то, что описано в: [K. П. Беннетт и О. Л. Мангасарян: "Различение двух линейно неразрывных множеств при устойчивом линейном программировании", Оптимизационные методы и программное обеспечение 1, 1992, 23-34].

Информация об атрибутах:

1) идентификационный номер

2) Диагноз (M = злокачественный, B = доброкачественный)

3-32)

Для каждого ядра клетки вычисляются десять характеристик с действительным знаком:

а) радиус (среднее расстояние от центра до точек по периметру)

б) текстура (стандартное отклонение значений шкалы серого)

в) периметр

г) площадь

д) гладкость (локальное изменение длины радиуса)

е) компактность (периметр ^ 2 / площадь - 1,0)

г) вогнутость (выраженность вогнутых участков контура)

з) вогнутые точки (количество вогнутых участков контура)

i) симметрия

к) фрактальная размерность («приближение береговой линии» - 1)

Среднее значение, стандартная ошибка и "худшее" или наибольшее (среднее из трех

наибольшие значения) этих характеристик были вычислены для каждого изображения,

в результате получается 30 функций. Например, поле 3 - это средний радиус, поле

13 - это радиус SE, поле 23 - наихудший радиус.

Все значения функций перекодированы с использованием четырех значащих цифр.

Отсутствующие значения атрибутов: нет

Распределение по классам: 357 доброкачественных, 212 злокачественных

4.Алгоритм Mapper. Описание алгоритма, созданного при помощи библиотеки keppler.

Mapper предоставляет краткое описание формы набора данных

(выраженное через кодомен отображения)

• Полезность Mapper заключается в его универсальности:

• Можно использовать любую функцию отображения

• Крышка может быть сконструирована произвольно

• Может быть использован любой алгоритм кластеризации

• Результирующий график часто гораздо проще интерпретировать, чем, например

, отдельные точечные диаграммы попарных взаимосвязей

• Mapper часто работает в паре с данными высокой размерности и

обычно используется для просмотра "истинной" формы или структуры данных.

• Mapper - это основной алгоритм, лежащий в основе компании AI, Ayasdi Inc.

• Борьба с отмыванием денег

• Выявление Мошенничества с Платежами

• Оценка рисков для здоровья

Простейшая интерпретация:

• Интерпретирует любой набор данных как

“данные облака точек”, превращает данные в

упрощенный топологический график

• “Mapper принимает в качестве входных данных как

возможно многомерный набор

данных, так и карту, определенную на основе данных, и

создает сводку данных, используя обложку кодомена карты".

Кодомен - множество функции, описанной символически как f : X -> Y- это множество Y, внутрь которого попадают все значения функции.

Для начала мы визуализируем результирующий график с помощью функции color_function, которая связывает с данными объектива их расстояние по x-координате с min, и цветовое сопоставление этих координат с заданной шкалой цветов Plotly. Здесь мы используем цветовую шкалу пивовара с шестнадцатеричным цветовым кодом.

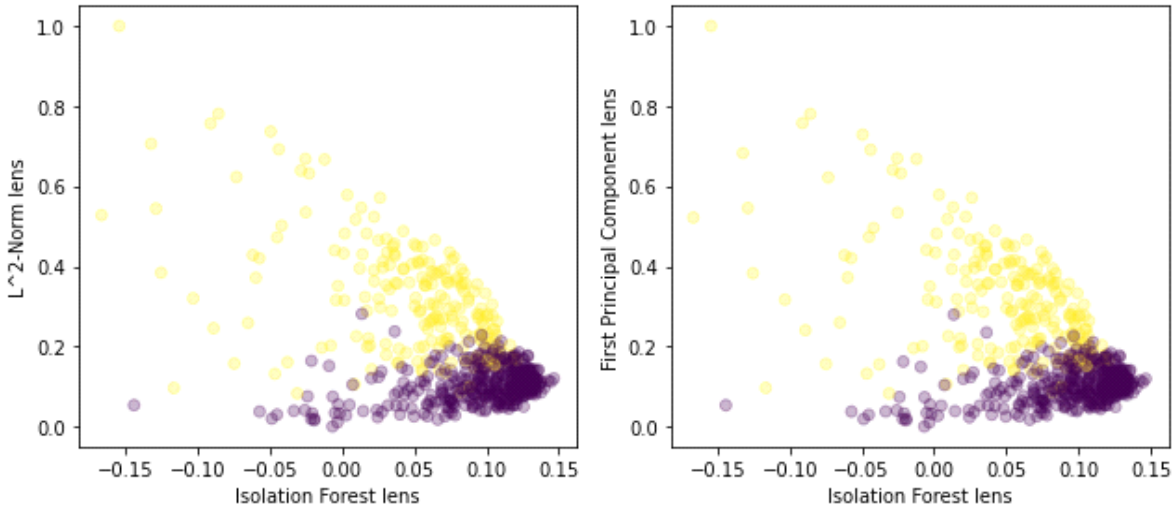
Причина выбора линз в демонстрации выше:

Для линзы 1: линзы, имеющие биологический смысл; Другими словами, линзы, которые выделяют особенности данных, о которых я знаю. В случае этих конкретных данных использование показателя аномалии (в данном случае рассчитанного с использованием IsolationForest из sklearn) имеет биологический смысл, поскольку раковые клетки являются аномальными.

Для линзы 2: линзы, которые рассеивают данные, а не объединяют множество точек вместе.

Ниже мы демонстрируем, что тот же результат может быть получен с другим выбором для линзы 2:

Прежде чем строить симплициальные комплексы с этими линзами, давайте исследуем влияние линз на данные:



Это показывает, что использование нормы L ^ 2 и первого главного компонента в качестве второй линзы, вероятно, сгенерирует очень похожие топологические графы.

Это также показывает, что первым основным компонентом данных является норма L ^ 2 или очень близкая к норме L ^ 2, и это может быть подтверждено путем проверки значений в массивах.

Рассмотрим код, написанный на языке Python.

```
df = pd.read_csv("data.csv")
feature_names = [c for c in df.columns if c not in ["id", "diagnosis"]]
df["diagnosis"] = df["diagnosis"].apply(lambda x: 1 if x == "M" else 0)
X = np.array(df[feature_names].fillna(0))
y = np.array(df["diagnosis"])
```

Рисунок 15.

Для данных мы используем датасет о раке молочной железы штата Висконсин. Считываем данные из таблицы формата csv. Создаем numpy-массивы для хранения полученных из таблицы данных.

```
model = ensemble.IsolationForest(random_state=1729)

model.fit(X)
lens1 = model.decision_function(X).reshape((X.shape[0], 1))
```

Создаем индивидуальную одномерную линзу с помощью изоляционного леса. Метод sklearn.ensemble.IsolationForest() возращает оценку аномалии каждой выборки с помощью алгоритма IsolationForest. Здесь параметр random_state управляет псевдослучайностью выбора признаков и значений разделения для каждого шага ветвления и каждого дерева в лесу.

```
mapper = km.KeplerMapper(verbose=3)
lens2 = mapper.fit_transform(X, projection="l2norm")


lens = np.c_[lens1, lens2]
```

Создаем еще одну одномерую линзу с L2-нормой. Объединим обе линзы, чтобы создать двумерную линзу: изоляционный лес и L^2-норму.

```
graph = mapper.map(
    lens,
    X,
    cover=km.Cover(n_cubes=15, perc_overlap=0.4),
    clusterer=sklearn.cluster.KMeans(n_clusters=2, random_state=1618033)
)
```

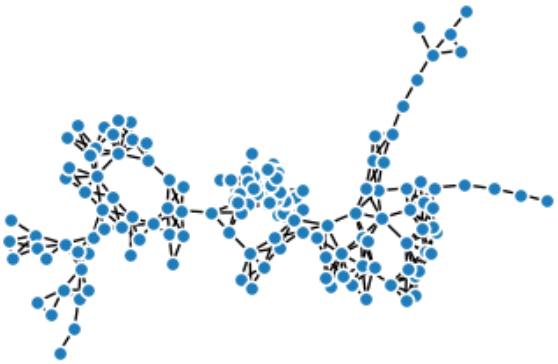
Данный код отвечает за построение графа. Метод sklearn.cluster.KMeans() за кластеризацию в ходе создания данного графа с помощью метода K-средних.

Остальная часть кода отвечает за различные способы визуализации

```
mapper.visualize(graph,
    path_html="breast-cancer.html",
    title="Wisconsin Breast Cancer Dataset",
    custom_tooltips=y
)
```

5.Анализ полученных данных.

В ходе работы алгоритма мы получили данный граф.



Баркод персистентного модуля — это просто мультимножество интервалов, которые используются в его разложении.

n-ые персистентные гомологии набора данных X — это персистентный модуль $V_a = H_n(C_a)$ с очевидными структурными отображениями. Его баркод и диаграмма дают некоторую визуальную информацию о структуре данных.

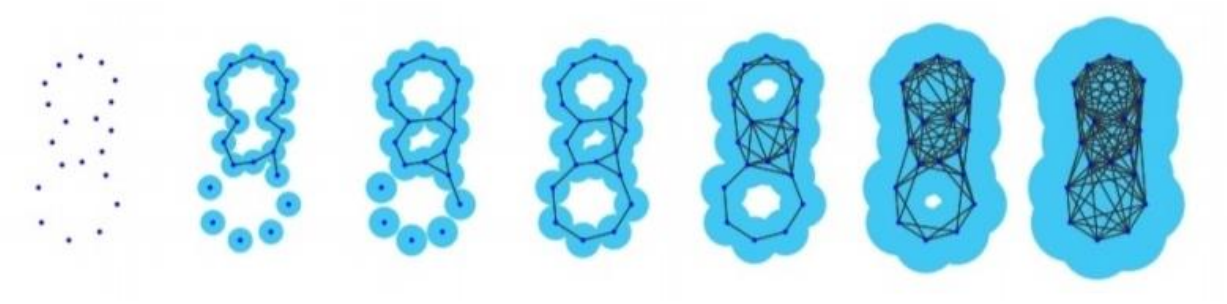
Связи между вершинами данного графа, обычно, в топологическом анализе данных не имеют большой важности, и не являются предметом исследования, т.к. такие характеристики, как расстояние между двумя вершинами, вообще говоря, крайне неустойчивы к шуму(помехам).Обычно облако, состоящее из набора точек, анализируют с помощью персистентных гомологий. Для этого рассматривают абстрактный симплициальный

комплекс C(X, E). Его вершинами являются элементы множества X, и $\{x_0, x_1, x_2,\ldots,x_n\}$, принадлежащее X, является симплексом, если пересечение замкнутых шаров радиуса E не является пустым:

$$B_E(x_0) \cap B_E(x_1) \cap \ldots \cap B_E(x_n) \neq \emptyset$$

Данный абстрактный комплекс называется комплексом Чеха

а)



б)

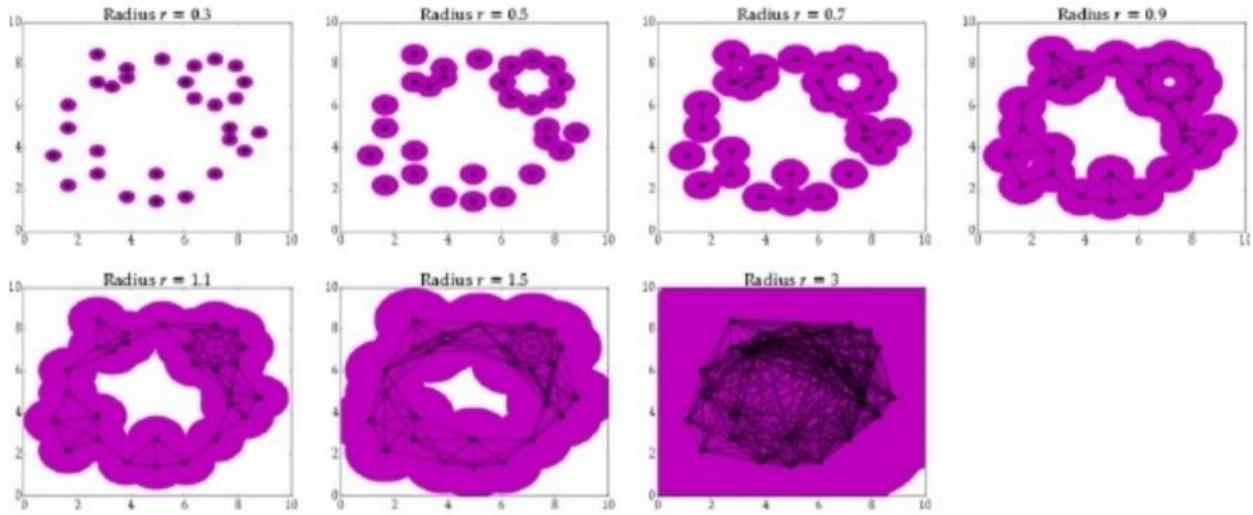


Рисунок 16.а,б. Пересечение шаров с параметризированным радиусом

Комплекс Чеха помогает получить топологическую информацию об облаке точек или распределении, потому что:

$H_n(C(X, E)) = H_n(X_r)$, где X_r – объединение шаров радиуса r с центрами в точках множества X.

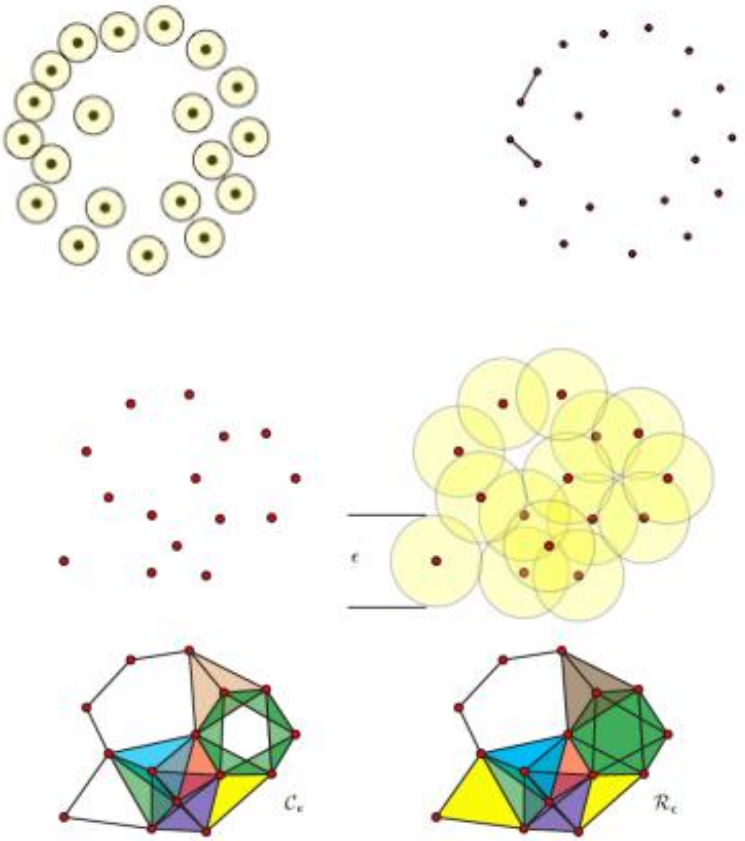


Рисунок 17. Процесс создание симплициального комплекса по облаку точек.

На рисунке 17. Видно, построение комплекса Чеха с разными радиусами окружностей.

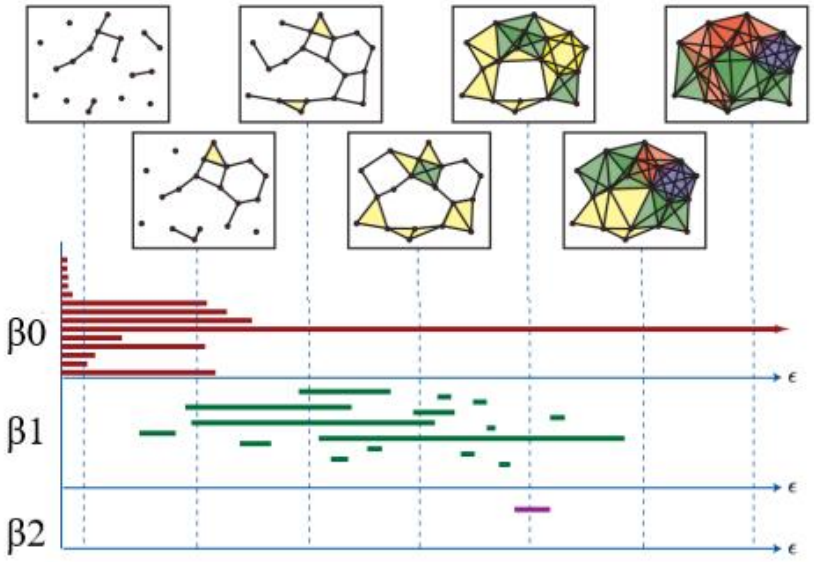


Рисунок 18.

На рисунке 18 рассматривается построение комплекса Чеха, а так же построение баркода по данному комплексу. Построено три баркода, каждый из них отвечает нулевому, первому и второму числам Бетти.

Нулевое число Бетти отвечает за количество компонент связности.Оно отвечает за представление “сливаемости кластеров” при увеличении радиуса окружностей.

Первое число Бетти интуитивно представляет собой максимальное число разрезов этого пространства, которые можно сделать без увеличения числа компонент связности. Оно указывает на 1-мерные “дыры”.

Однозначной интерпретации получаемой таким образом информации не существует. Дата-саентисты и ученые, специализирующиеся на конкретной научной области, изучают топологическую структуру, “Невозможные позиции” и т.д.

Однако иногда Mapper помогает обнаружить закономерности или новые данные без использования топологического аппарата.

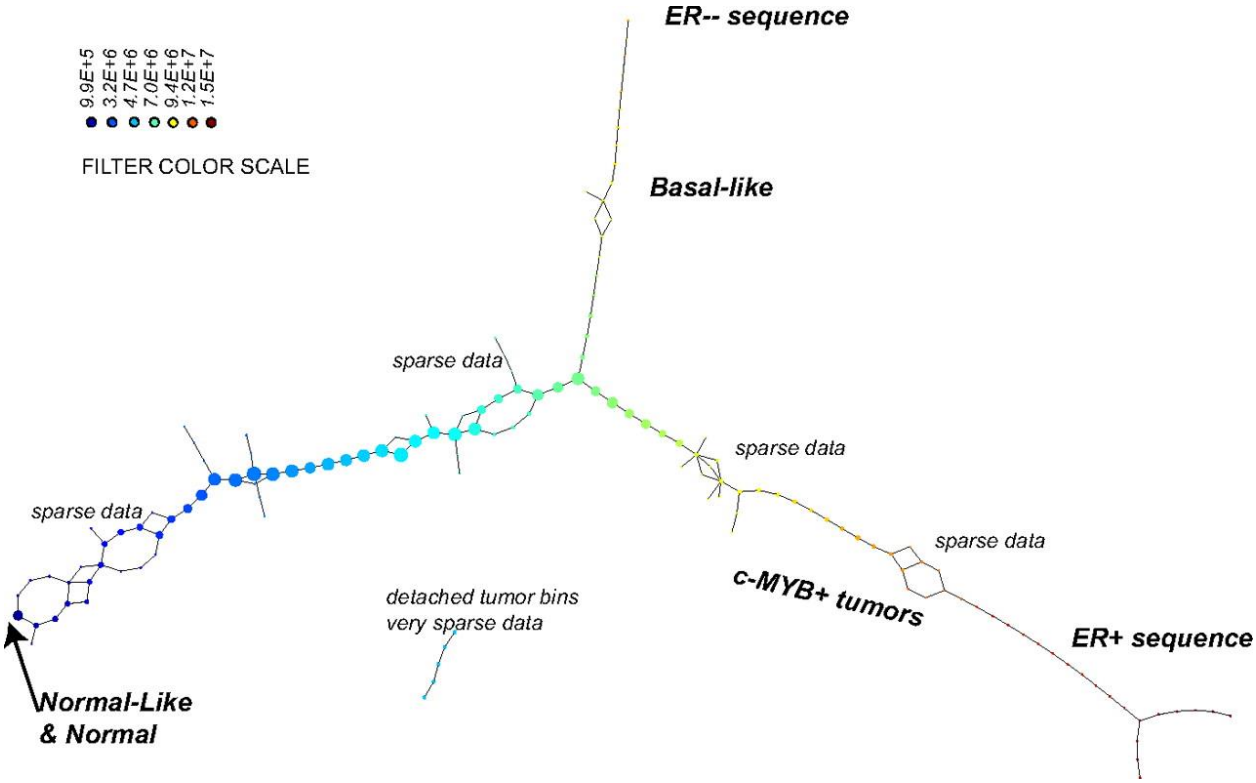
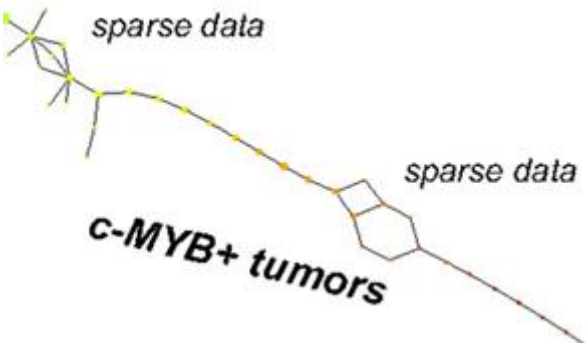


Рисунок 19. Граф, созданный с помощью алгоритма Mapper при исследовании рака молочной железы.

Как видно, данный граф имеет три ответвления, в них скоплены три различные группы данных: которые отвечают эстроген-рецептор

положительным (ER^+), эстроген-рецептор отрицательным(ER^-) и ракам нормального типа(Normal-like & Normal). Оказалось, что в ER^+ ответвлении есть участок, необычная форма которого определяется такой характеристикой, как высокая экспрессия c-MYB-гена. Его назвали $c\text{-}MYB^+$.



Оказалось, что рак данного типа имеет 100% выживаемость в случае метастазирования, несмотря на то, что само ответвление(ER^+) является красным, то есть крайне опасным для здоровья человека, а так же крайне сложно поддается лечению.

Заключение

Входе данной работы я изучил несколько исследований мировых ученых, в которых важную роль играли инструменты топологического анализа данных, основы математического аппарата, применяемого в данной области data science. Я изучил возможности алгоритма Mapper на наборе данных о пациентах больницы Висконсина.

Топологический анализ данных – область data science, которая поможет найти закономерности во многих областях науки и станет одним из основных и популярных направлений data science.

Список использованной литературы.

https://ru.ert.wiki/wiki/Topological_data_analysis#Characteristics_of_TDA_in_applications

https://www.researchgate.net/publication/305645903_Persistent_Homology_for_Fast_Tumor_Segmentation_in_Whole_Slide_Histology_Images

https://doi.org/10.1016/j.procs.2016.07.033

https://royalsocietypublishing.org/doi/10.1098/rsif.2014.0873

https://www.youtube.com/playlist?list=PLDWwKK3tSCltvFQ2Cypp0vCJSy0XdiaPH

python scikit-TDA: http://espressocode.top/hand-written-digits-using-topological-data-analysis/

https://colorbrewer2.org/#type=sequential&scheme=OrRd&n=3

https://kepler-mapper.scikit-tda.org/en/latest/generated/gallery/plot_breast_cancer.html#sphx-glr-download-generated-gallery-plot-breast-cancer-py

https://kepler-mapper.scikit-tda.org/en/latest/notebooks/Cancer-demo.html

https://kepler-mapper.scikit-tda.org/en/latest/generated/gallery/plot_breast_cancer.html#sphx-glr-generated-gallery-plot-breast-cancer-py

А.Б. Скопенков “Алгебраическая топология с геометрической точки зрения”

А.И. Кострикин “Введение в алгебру”

М. Э. Казарян “Введение в теорию гомологий”

Raul Rabadan “Topological Data Analysis for Genomics and Evolution”