Diffusion maps

Елизавета Кияко/Фуад Бабаев

AI MASTERS

kiyako_2002@mail.ru / f.babaev@yahoo.com
GitHub

25 декабря 2023 г.

Введение

1. **"О чём"**
   В данной презентации будет представлена концепция диффузионных карт и их применение в анализе данных.
2. **"Зачем"**
   С ростом объёмов данных и увеличением сложности структур важно иметь инструменты для выявления внутренних связей в данных, что позволяет более эффективно проводить анализ и визуализацию.
3. **"Гипотеза"**
   Предполагается, что использование диффузионных карт обеспечит более глубокое понимание структуры данных и выявит скрытые шаблоны, недоступные при применении традиционных(линейных) методов анализа.

Применение

1. Задача снижения размерности
2. Задача кластеризации
3. Задача детекции выбросов

Основные идеи метода

Algorithm Базовый алгоритм создания диффузионной карты

Require: $X_i, i = 0 \dots N - 1.$

1. Определить ядро $k(x, y)$ и создать матрицу ядра K, такую что $K_{ij} = k(X_i, X_j)$.
2. Создать матрицу диффузии, нормализовав строки матрицы ядра.
3. Вычислить собственные векторы матрицы диффузии.
4. Отобразить в d-мерное диффузионное пространство за "время" t, используя d доминирующих собственных векторов и значений.

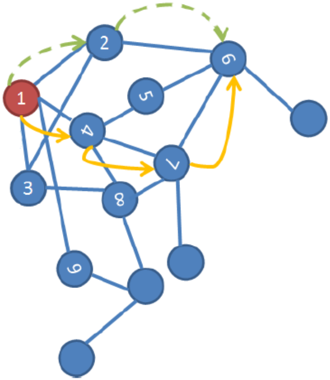
Output: Данные пониженной размерности $Y_i, i = 0 \dots N - 1.$
=0

Основные идеи метода

$$\text{connectivity}(x, y) = p(x, y). \quad (1)$$

$$\text{connectivity}(x, y) \propto k(x, y). \quad (2)$$

$$k(x, y) = \exp\left(-\frac{|x - y|^2}{\alpha}\right). \quad (3)$$



Переходы на графе

Основные идеи метода

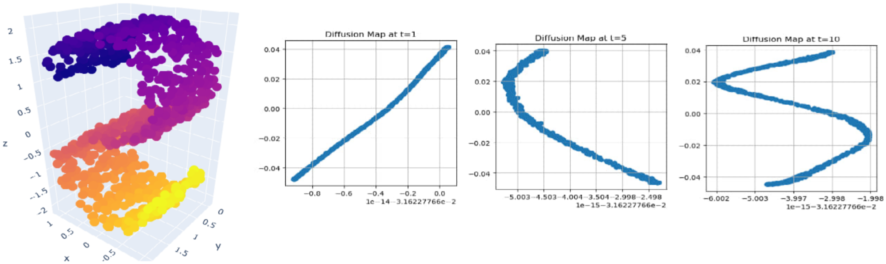
$$\frac{1}{d_x}\sum_{y\in X} k(x,y) = 1. \tag{4}$$

$$\text{connectivity}(x,y) = p(x,y) = \frac{1}{d_x} k(x,y) \tag{5}$$

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

$$P^2 = \begin{bmatrix} p_{11}p_{11} + p_{12}p_{21} & p_{12}p_{22} + p_{11}p_{12} \\ p_{21}p_{11} + p_{22}p_{21} & p_{22}p_{22} + p_{21}p_{12} \end{bmatrix}$$

Пример калибровки параметра t



Калибровка параметра t

Ключевые особенности

$$Y_i := \begin{bmatrix} p_t(X_i, X_1) \\ p_t(X_i, X_2) \\ \vdots \\ p_t(X_i, X_N) \end{bmatrix} = P_{i*}^T \tag{6}$$

$$Y_i' = \begin{bmatrix} \lambda_1^t \psi_1(i) \\ \lambda_2^t \psi_2(i) \\ \vdots \\ \lambda_n^t \psi_n(i) \end{bmatrix} \tag{6}$$

$\psi_1(i)$ характеризует i-й элемент первого собственного вектора матрицы P.

Вычисление собственных значений больших разреженных матриц

Для вычисления собственных значений использовался метод **Implicitly Restarted Arnoldi method (IRAM)**

- Используя метод Арнольди, строится базис Крыловского подпространства и формируется верхняя эрмитова матрица Гессенберга.
- С помощью спектрального сдвига (shift) выделяются желаемые собственные значения, которые "сдвигаются"к началу списка собственных значений матрицы Гессенберга.
- Применяется неявный перезапуск, который модифицирует базис подпространства таким образом, чтобы избавиться от влияния отвергнутых собственных значений, сохраняя при этом желаемые собственные значения.
- Итерационный процесс продолжается до тех пор, пока не будет достигнута желаемая точность или не будут найдены все требуемые собственные значения.

Оценка сложности

DM

$O(n^3)$ - буквально сложность IRAM.

PCA

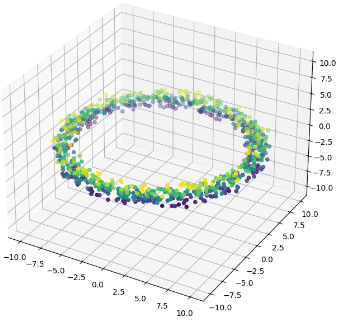
$O(n^3)$

t-SNE

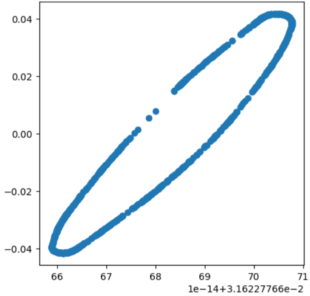
$O(n^2 \log n)$

Снижение размерности данных. Тор

Сгенерируем данные, имеющие структуру незашумленного тора, и посмотрим на то, как диффузионные карты отразят его в двумерное пространство.



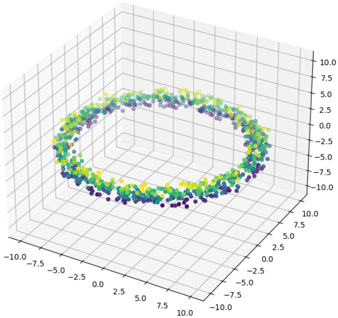
Незашумленный тор



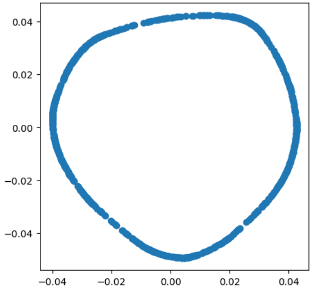
Первая реализация снижения размерности через диффузионные карты

Снижение размерности данныхю. Тор

Уберем собственный вектор, соответсвующий собственному числу $\lambda = 1$.



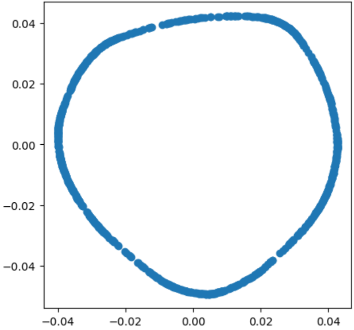
Незашумленный тор



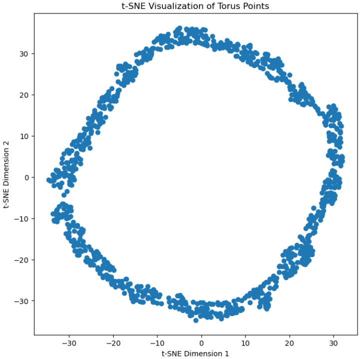
Вторая реализация снижения размерности через диффузионные карты

Снижение размерности данных. Тор. Сравнение с TSNE

Сравним нашу реализацию с алгоритмом TSNE.



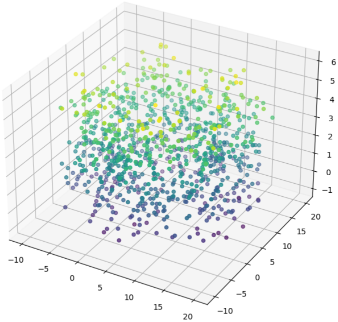
Реализация снижения размерности через диффузорные карты



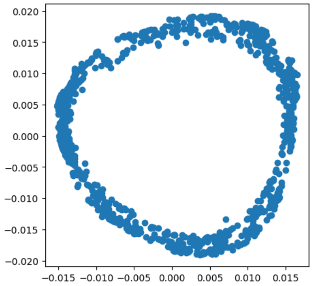
TSNE

Снижение размерности данных. Тор. Устойчивость

Зашумим тор и построим отображение в двумерное пространство.



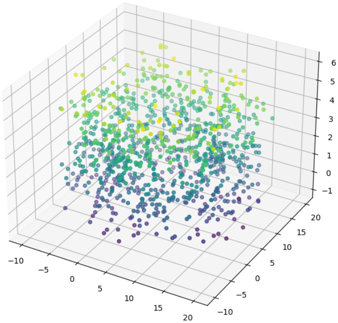
Зашумленный тор



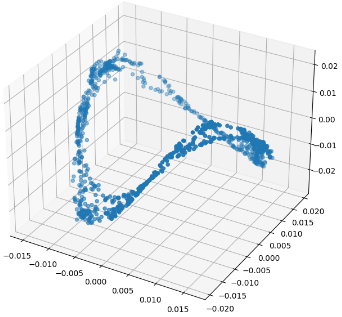
Реализация снижения размерности с помощью диффузорных карт

Снижение размерности. Тор

Для понимания принципа работы диффузионных карт, построим с их помощью отображение в трехмерное пространство.



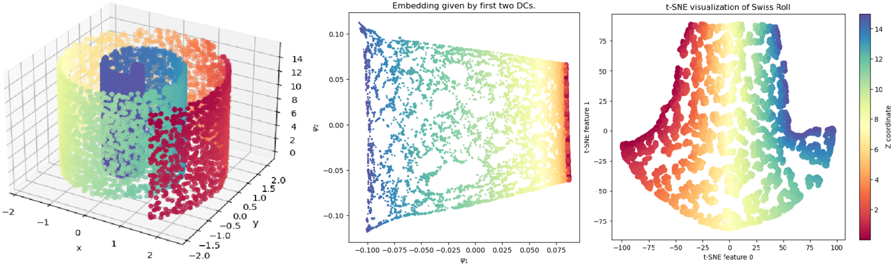
Зашумленный тор



Отображние в трехмерное пространство

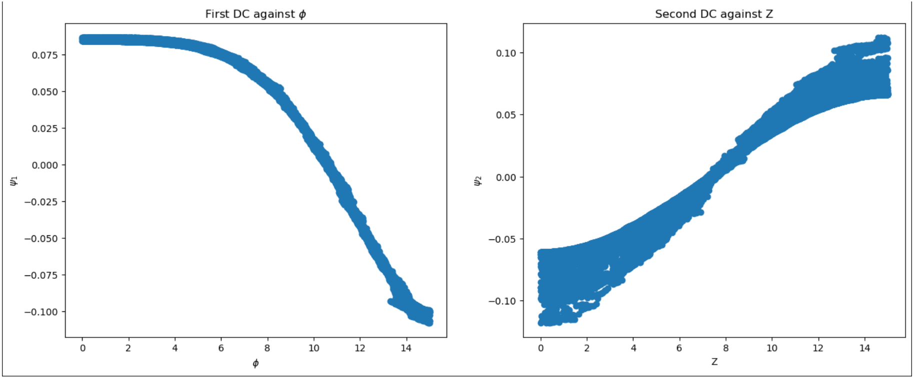
Увидим, что геометрия данных передается корректно.

Снижение размерности. Swiss roll (рулет)



Снижение размерности для swiss roll

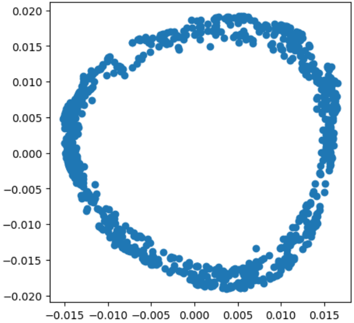
Снижение размерности. Swiss roll (рулет)



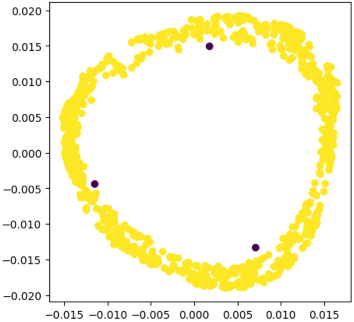
Оценка корреляции собственных значений с параметрами рулета

Детектирование аномалий. 2D

Детекцию аномалий можно привести, например, с помощью алгоритма DBSCAN. Темно фиолетовые точки на картинке ниже означают, что точка выбивается из общего паттерна данных

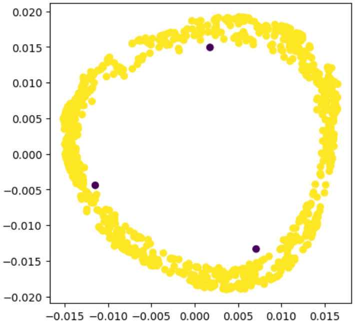


Реализация снижения размерности тора с помощью диффузорных карт

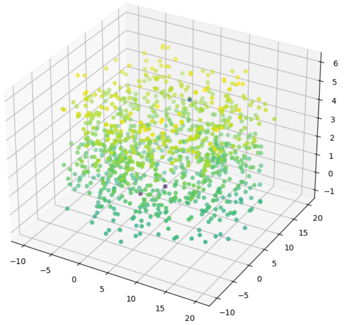


Выявленные аномалии

Посмотрим на прообразы аномалий, которые мы нашли в двумерном пространстве, в трехмерии. Видно, что в исходных трехмерных данных мы нашли только самые очевидные аутлайеры.

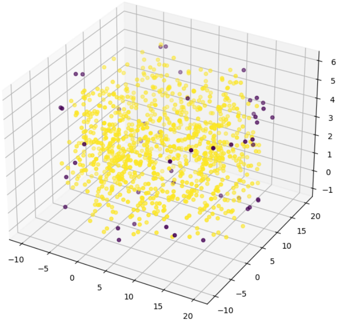


Выявленные в 2D аномалии

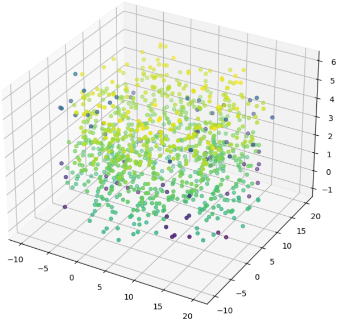


Выявленные в 3D аномалии

Детектирование аномалий. 3D



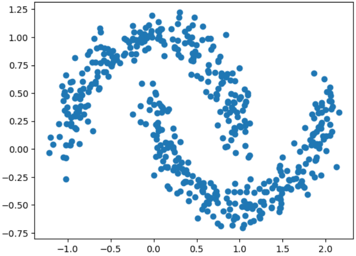
Выявленные аномалии в 3D для большого ε



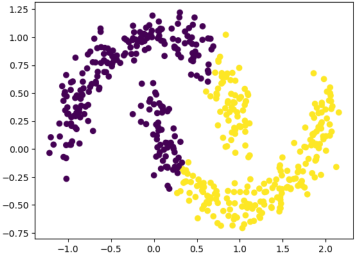
Выявленные аномалии в 3D для малого ε

Кластеризация

Сгенерируем данные в виде двух лун. Посмотрим, как справится с их кластеризацией алгоритм KMeans.



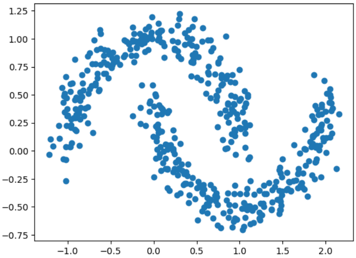
Данные



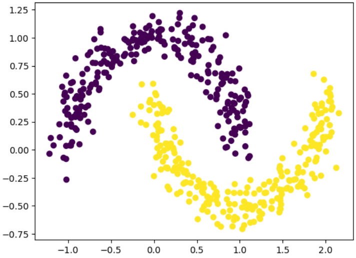
Кластеризация с помощью kmeans

Кластеризация

Сравним с кластеризацией с помощью диффузионных карт.



Данные



Кластеризация с помощью диффузорных карт

Основные результаты проекта

1. Были реализованы алгоритмы снижения размерности, кластеризации и детекции аномалий в данных с помощью диффузионных карт.
2. Алгоритмы были применены к синтетическим данным для проверки их базовой работоспособности. Так же были приведены некоторые сравнения с иными алгоритмами для решения поставленных задач.
3. Диффузионные карты вкладывают исходные данные в пространства большей или меньшей размерности, сохраняя их топологическую структуру, что может быть полезно для решения совершенно разных задач (не только тех, которые рассмотрели мы)

Заключение

1. **"Что планировалось"**
   Планировалось реализовать методы снижения размерности, кластеризации и детекции аномалий в данных с помощью диффузионных карт. А так же применить эти методы к различным данным.
2. **"Что получилось, а что нет"**
   Получилось реализовать рещения для всех трех поставленных задач с помощью диффузионных карт, опробовать их работу на синтетических данных, а так же провести некоторые сравнения с результатами работы иных алгоритмов.
   Не удалось применить эти методы к реальным данным, например к MNIST датасету.

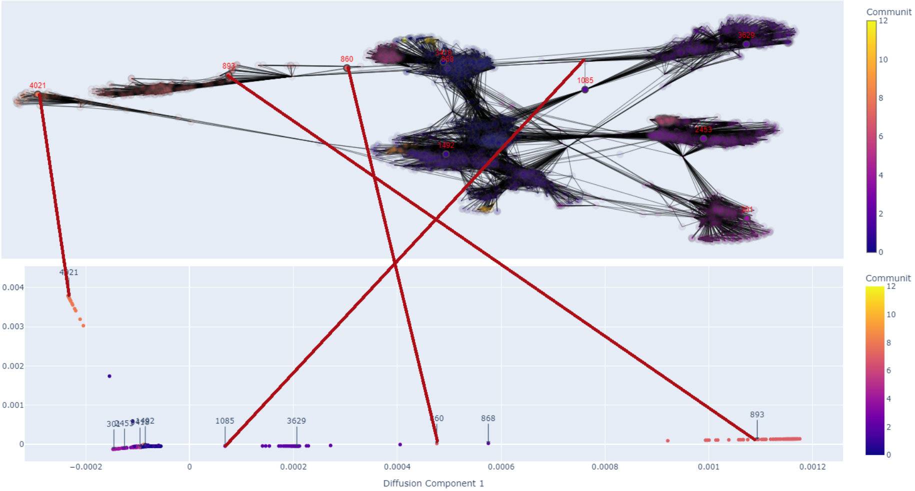
References

Coifman, R. R., & Lafon, S. (2006).
Diffusion maps.
Applied and Computational Harmonic Analysis, 21(1), 5–30.
doi: 10.1016/j.acha.2006.04.006.

Nadler, B., Lafon, S., Coifman, R., & Kevrekidis, I. G. (2008).
Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms.
In *Principal Manifolds for Data Visualization and Dimension Reduction*, 238–260.
doi: 10.1007/978-3-540-73750-6_10.

de la Porte, J., Herbst, B. M., Hereman, W., & van der Walt, S. J. (2009).
An Introduction to Diffusion Maps.
Applied Mathematics Division, Department of Mathematical Sciences, University of Stellenbosch, South Africa; Colorado School of Mines, United States of America.

Appendix



Применение DM на Stanford Large Network (Social circles from Facebook)