Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Физтех-школа прикладной математики и информатики
Кафедра дискретной математики

Выпускная квалификационная работа магистра

Топологические методы в некоторых задачах анализа данных

Автор:
Студент 115 группы
Снопов Павел Михайлович

Научный руководитель:
д-р физ.-мат. наук, проф.
Мусин Олег Рустумович

Москва 2023

Аннотация

Топологические методы в некоторых задачах анализа данных
Снопов Павел Михайлович

Краткое описание задачи и основных результатов, мотивирующее прочитать весь текст

Abstract

Topological methods in some problems of data analysis

Оглавление

Глава 1

Введение

В этой части надо описать предметную область, задачу из которой вы будете решать, объяснить её актуальность (почему надо что-то делать сейчас?). Здесь же стоит ввести определения понятий, которые вам понадобятся в постановке задачи.

Глава 2

Необходимые теоретические сведения

В этой главе мы напомним некоторые определения и факты, используемые в дальнейшем.

2.1 Симплициальные гомологии

Пусть K – симплициальный комплекс, то есть множество симплексов, такое, что

1. Для каждого симплекса из K его грани тоже лежат в K.
2. Пересечение любых двух симплексов $\sigma, \tau \in K$ либо пусто, либо является гранью и σ, и τ.

Тогда можно рассмотреть *n-мерные цепи c*, т.е. линейные комбинации с целыми коэффициентами (лишь конечное число которых ненулевые) всех ориентированных n-симплексов в K

$$c = \sum_i z_i \sigma_i^n.$$

Множество $C_n(K)$ всех n-цепей называется *n-й группой цепей* и имеет очевидную структуру свободного $\mathbb{Z}$-модуля. Прямая сумма $\bigoplus_{n\geq 0} C_n(K)$ таких групп также является свободным $\mathbb{Z}$-модулем и обозначается $C_*(K)$.

Из n-цепи можно получить $(n-1)$-цепь путем взятия границы. А именно, *границей* n-цепи $c = \sum_i z_i \sigma_i^n$ называется $(n-1)$-цепь

$$\partial_n(c) = \sum_{j=0}^{n} (-1)^j \sum_i \varepsilon_i \partial_j \sigma_i^n,$$

где $\partial_j \sigma_i^n = \partial_j [v_0, ..., v_n] = [v_0, ..., \hat{v}_j, ..., v_n]$ – это $(n-1)$-симплекс, порожденный всеми вершинами, кроме вершины v_j. Гомоморфизм модулей ∂_n называется *граничным оператором.*

Лемма (Пуанкаре)**.** *Для любого $n \geq 2$ справедливо*

$$\partial_{n-1} \circ \partial_n = 0.$$

Последовательность $\mathbb{Z}$-модулей и гомоморфизмов ∂ между ними (на самом деле, вместо $\mathbb{Z}$-модулей и гомоморфизмов могут быть объекты любой аддитивной категории и морфизмы между ними [1]), для которых верно, что $\partial^2 = 0$, называется *цепным комплексом.*

Нетрудно заметить, что в таком случае $\operatorname{im} \partial_{n+1} \leq \ker \partial_n$.

Определение 1 (Группа гомологий)**.** *n-й группой гомологий $H_n(K)$ симплициального комплекса K называется $\mathbb{Z}$-модуль*

$$H_n(K) = {\ker \partial_n} \big/ {\operatorname{im} \partial_{n+1}}.$$

2.2 Pas de Deux: категории и функторы

[2] Конструкция групп гомологий обладает одним замечательным свойством: она функториальна. Для того, чтобы наиболее полно описать это, придется воспользоваться аппаратом теории категорией.

Категорией $\mathbf{C}$ называют коллекцию объектов и морфизмов между парой объектов. То есть, категория состоит из

1. (классов) объектов X,Y,Z,
2. для любых двух объектов X и Y существует класс морфизмов $\mathrm{Hom}_{\mathbf{C}}(X,Y)$ (если категория понятна из контекста, то индекс будем опускать),
3. для любого объекта X существует единичиный морфизм $1_X \in \mathrm{Hom}(X,X)$,
4. на морфизмах задана операция композиции: для любого пары морфизмов $f \in \mathrm{Hom}(X,Y)$ и $g \in \mathrm{Hom}(Y,Z)$ существует морфизм $gf \in \mathrm{Hom}(X,Z)$,
5. для любого морфизма $f \in \mathrm{Hom}(X,Y)$ верно, что $1_Y f = f = f 1_X$,
6. операция композиции ассоциативна.

......

Если имеется два симплициальных комплекса K и L, и симплициальное отображение f между ними, то оно индуцирует гомоморфизм модулей

$$f_{*,n} : H_n(K) \to H_n(L)$$

2.3 Персистентные гомологии

[3], Под облаком точек D здесь и далее мы будем подразумевать конечное множество в $\mathbb{R}^n$. О персистентных (устойчивых) гомологиях можно думать как об адаптации понятия гомологии к облаку точек.

А именно, имея облако точек, можно построить семейство симплициальных комплексов, параметризованное некоторым частично упорядоченным множеством (например, на практике зачастую это $\mathbb{R}$ или его подмножества). Такая параметризация естественным образом задает отображения между получаемыми симплициальными комплексами. Тогда устойчивыми гомологиями будут в точности гомологии полученных комплексов вместе с индуцированными гомоморфизмами между ними.

Определим теперь устойчивые гомологии формально.

Зафиксируем некоторое частично упорядоченное множество $T \subseteq \mathbb{R}$ и соответствующую ему категорию $\mathbf{T}$. Тогда функтор

$$F : \mathbf{T} \to \mathbf{Simp}$$

называется *фильтрацией.*

Наиболее важным для нас примером фильтрации на облаке точек D будет являться *фильтрация Вьеториса-Рипса*

$$\mathrm{Rips}(D) : [\mathbf{0},\infty) \to \mathbf{Simp} : r \mapsto X(\mathcal{N}(D)_r),$$

где $\mathcal{N}(D)_r$ – это граф соседей радиуса r, т.е. такой граф, вершины которого –это исходное облако точек D, и ребро существует, если расстояние между двумя точками не превосходит r. $X(G)$ – это кликовый комплекс графа G, т.е. наибольший (по включению) симплициальный комплекс, чей 1-скелет – это G.

Алгоритм 1: Алгоритм построения комплекса Вьеториса—Рипса

Исходные параметры: Облако точек X, вещественное число $\alpha > 0$.
Результат: Симплициальный комплекс Вьеториса—Рипса
Для каждой точки x строим её α-окрестность $B_\alpha(x)$;
$i = 1$;
до тех пор, пока $i + 1$ *окрестностей попарно имеют непустое пересечение* **выполнять**
  строим i-ый симплекс на соответствующих вершинах;
  $i \leftarrow i + 1$;
конец

Композиция функтора гомологий H_n : $\mathbf{Simp} \to \mathbb{Z}{-}\mathbf{Mod}$ с фильтрацией даст функтор

$$M_n : \mathbf{T} \to \mathbb{Z}{-}\mathbf{Mod}.$$

Этот функтор называется *n-ым устойчивым модулем.* Совокупность таких модулей по n, варьирующееся по всем размерностям, называется *устойчивыми гомологиями* $PH_*(T)$ фильтрации T.

На практике часто используют фильтрацию Вьеториса-Рипса, тогда весь метод выглядит так: имея облако точек D, варьируя радиус r, строится возрастающая последовательность симплициальных комплексов

$$K_0 \subseteq K_1 \subseteq ... \subseteq K_s \subseteq ...,$$

где каждый симплициальный комплекс K_j является кликовым комплексом графа $\mathcal{N}(D)_j$ соседей радиуса j. Применяя к полученной последовательности функтор гомологий, получается последовательность $\mathbb{Z}$-модулей

$$H_n(K_0) \to H_n(K_1) \to ... \to H_n(K_s) \to ....$$

Это и есть устойчивые гомологии.

Из такого определения сразу видно, что устойчивые модули – это просто представление частично упорядоченного множества $(T, \leq)$. В хороших случаях имеется целая плеяда утверждений, которые описывают структуру таких модулей. Ниже представлено одно из наиболее общих, и одновременно наиболее подходящих для наших задач, утверждений.

Теорема 1 (Crawley-Boevey[4])**.** *Пусть k – поле, и $T \subseteq \mathbb{R}$. Тогда любое поточечно-конечное представление V частично упорядоченного множества $(T, \leq)$ над k имеет вид*

$$V \simeq \bigoplus_{I \in \mathcal{B}(V)} k_I,$$

где I – некоторый интервал $[b,d]$ в T, а k_I – это представление, которое имеет вид

$$... \to \underset{b-1}{0} \to \underset{b}{k} \xrightarrow{\mathrm{id}} ... \to \underset{d}{k} \xrightarrow{\mathrm{id}} \underset{d+1}{0} \to ...$$

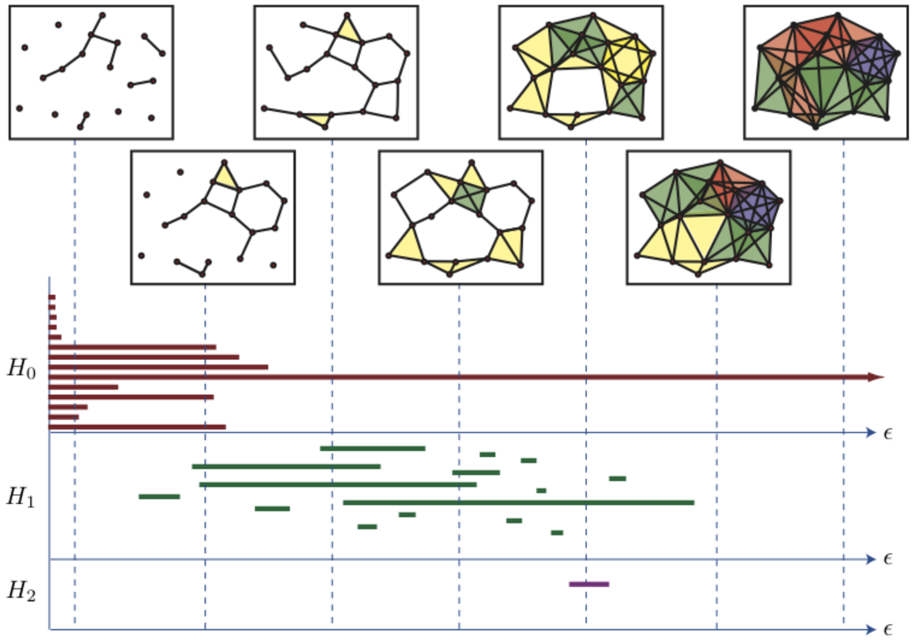

Рис. 2.1: Пример фильтрации и полученного по ней баркода

Множество $\mathcal{B}(V)$ называется *баркодом* данного устойчивого модуля. Каждый интервал в баркоде содержит в себе информацию о том, когда определенный топологический признак появился – время рождения, и когда исчез – время смерти.

Эту же информацию можно представлять иначе: каждый интервал $[b,d]$ баркода можно рассматривать как точку с координатами (b,d) на (расширенной) плоскости. Так как $0 < b < d$, то эти точки всегда находятся в положительном квадранте выше диагонали $y = x$. Такое представление баркода называется *диаграммой устойчивости*.

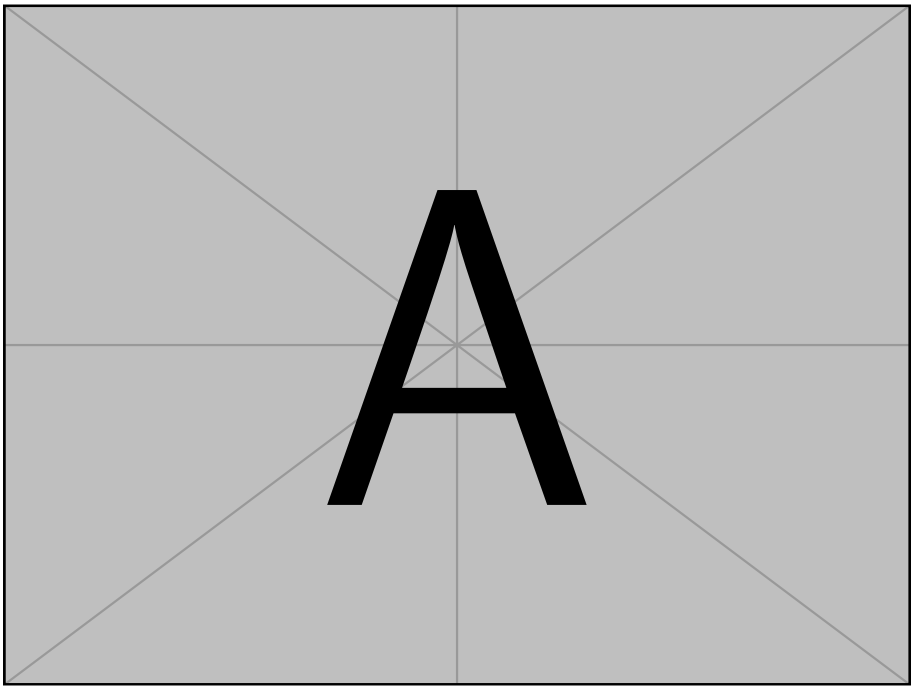

Рис. 2.2: Пример диаграммы устойчивости

Глава 3

Разработка алгоритма симплификации временного ряда на основе устойчивых гомологий

3.1 Алгоритм Рамера-Дугласа-Пекера

Рассмотрим временной ряд X, т.е. последовательность точек, наблюдений, индексированных временем наблюдения. Часто возникает задача упрощения получаемой кривой, например, для сглаживания временного ряда. Стандартным методом решения такой задачи является алгоритм Рамера-Дугласа-Пекера. Это алгоритм, основанный на построении ломаной линии, которая аппроксимировала бы исходный временной ряд, но при этом содержала бы меньшее число точек.

Основной сложностью алгоритма является поиск подходящего гиперпараметра ε.

Алгоритм Рамерка-Дугласа-Пекера, в силу своей конструкции, сохраняет минимумы и максимумы. Такого же эффекта можно добиться при помощи устойчивых гомологий, правильно подобрав фильтрацию. Рассмотрим это.

3.2 Симплификация на основе устойчивых гомологий

Рассмотрим временной ряд $X : \mathbb{N} \to \mathbb{R}$. Его график можно воспринимать как 1-сиплициальный комплекс, где 1-симплексами будут отрезки, соединяющие соседние точки временного ряда. Тогда можно рассмотреть следующую фильтрацию

$$F : \mathbf{R} \to \mathbf{Simp} : r \mapsto X^{-1}(-\infty, r).$$

По данной фильтрации можно посчитать устойчивые гомологии. В силу конструкции симплициального комплекса, только 0-мерные устойчивые гомологии будут нетривиальны. Более того, точки экстремума такого отображения будут отвечать за рождение и смерть 0-мерных циклов (что следует общему механизму, развитому в теории Морса, см. [**MorseTheory**]). Полученные в результате подсчета точки на диаграмме устойчивости будут кодировать в себе как раз такие экстремальные точки. В качестве симплифицированного временного ряда можно рассмотреть все точки временного ряда, которые либо отвечали за рождение, либо за смерть 0-мерных циклов.

Можно рассмотреть параметризованную версию такого алгоритма, а именно ввести параметр $\varepsilon \geq 0$, отвечающий за расстояние до диагонали на диаграмме устойчивости. Тогда если расстояние точки на диаграмме меньше ε, то такая точка исключается из построения симплифицированного временного ряда.

3.3 Сравнение двух методов

Сравним алгоритм симплификации на основе устойчивых гомологий с алгоритмом Рамера-Дугласа-Пекера, используя l^2 метрику. Сравнение проведем на 2 различных областях: рассмотрим временные ряды, связанные с измерением температуры, а также временные ряды, отвечающие за ежедневную цену акций.

Глава 4

Описание практической части

Если в рамках работы писался какой-то код, здесь должно быть его описание: выбранный язык и библиотеки и мотивы выбора, архитектура, схема функционирования, теоретическая сложность алгоритма, характеристики функционирования (скорость/память).

Глава 5

Заключение

Здесь надо перечислить все результаты, полученные в ходе работы. Из текста должно быть понятно, в какой мере решена поставленная задача.

Список использованных источников

1. Weibel C. An Introduction to Homological Algebra. — Cambridge University Press, 1994. — (Cambridge Studies in Advanced Mathematics). — URL: https://books.google.ru/books?id=flm-dBXfZ%5C_gC.
2. Riehl E. Category Theory in Context. — Dover Publications, 2017. — (Aurora: Dover Modern Math Originals). — URL: https://books.google.ru/books?id=6B9MDgAAQBAJ.
3. Schenck H. Algebraic Foundations for Applied Topology and Data Analysis. — Springer International Publishing, 2022. — (Mathematics of Data). — URL: https://books.google.ru/books?id=4IQkzwEACAAJ.
4. Crawley-Boevey W. Decomposition of pointwise finite-dimensional persistence modules. — 2014.