УДК 004.67

© 2020 г. **О.Н. Артамонов**

(Дальневосточный федеральный университет, Владивосток)

ТОПОЛОГИЧЕСКИЕ МЕТОДЫ В ОБРАБОТКЕ И ИНТЕРПРЕТАЦИИ БОЛЬШИХ ДАННЫХ

В статье рассмотрено приложение методов алгебраической топологии к извлечению неочевидной полезной информации из больших массивов числовых данных. Это позволяет вычленять качественные характеристики объекта или явления, представляемого этими упорядоченными облаками чисел. В дальнейшем рассматриваются примеры практического использования полученной такими методами информации о многомерных дырах в данных.

Ключевые слова: математическое моделирование, аппроксимационный симплициальный комплекс, фильтрация комплексов, анализ данных, персистентные гомологии, числа Бетти, топологические инварианты, вычислительная топология, многомерная связность.

DOI: 10.22250/isu.2020.64.17-31

Введение

В настоящее время особенно актуально встал вопрос обработки больших данных. Компании и организации накапливают в процессе своей работы гигантские по объему и сложные по структуре массивы данных, для эффективного использования которых необходимо уметь их анализировать и интерпретировать. Если данные представляют собой неструктурированные текстовые материалы, то обработать их математическими средствами крайне сложно, скорее это поле деятельности для гуманитариев и архивариусов. Здесь нас будут интересовать только числовые или трансформируемые в числовой формат данные, при составлении которых были соблюдены единые принципы и системный подход.

Когда мы хотим обработать большой объем данных, то основная проблема состоит в том, как к ним подступиться. Как правило, оцифрованные

данные представляют собой гигантский числовой массив, и это облако чисел нуждается в упорядочении: подведении под единый формат, избавлении от “шума”, а порой даже в восстановлении утраченных или неполных данных. Желательно иметь возможность применить подходящие по характеру задачи математические техники. Для этого на основе упорядоченных данных нужно создать упрощенную математическую модель, которая бы отражала важные особенности исходной информации.

В этой статье показано, как упрощенная аппроксимационная модель позволяет вычислять информацию о связности исходных числовых данных. Эти данные для обработки, как правило, представляют собой статистику какого-либо процесса или явления, – например, множество параметров сигналов, пришедших на сеть расположенных на поверхности датчиков, статистика пациентов госпиталей и т.п. Мы представляем единичную информацию на входе как точку, чья размерность равна количеству качественных характеристик интересующих нас объектов или явлений. Пусть $X \subseteq \mathbb{R}$ будут входные данные для обработки, представляющие собой множество d -мерных точек, представленных в удобном для вычисления формате. На основе этого облака чисел возможно создать упрощенную математическую модель размерности d , сохраняющую, тем не менее, все топологические характеристики исходного облака чисел [1, 2].

Аффинно независимое множество $V \subseteq X \subseteq \mathbb{R}$, $V = \{v_0, v_1, \dots, v_k\}$ точек определяет k -симплекс $\sigma_V = \text{conv}(V) = \langle v_0, v_1, \dots, v_k \rangle$ размерности $k = \dim(\sigma_V) = \text{card}(V) - 1$, чьими вершинами являются точки $\langle v_i \rangle$, которые сами являются 0-симплексами; k -мерный симплекс σ_k , или просто k -симплекс, состоит из $(k + 1)$ -го элемента; σ_0 являются одномерными точками; σ_1 – отрезками; σ_2 – треугольниками; σ_3 – тетраэдрами и т.д.; σ_k – k -грань q -симплекса σ_q , если каждая вершина σ_k является также вершиной σ_q . Объединение граней симплекса σ_q составляет его границу.

Симплициальный комплекс представляет собой конечное семейство аккурратно, грань к грани, “склеенных” симплексов. Формально, симплициальный комплекс определяется как конечное семейство симплексов – такое, что:

- 1) если $\sigma \in K$ и τ является гранью $\sigma \Rightarrow \tau \in K$;
- 2) если $\sigma, \bar{\sigma} \in K \Rightarrow \sigma \cap \bar{\sigma}$ пусто или общая грань для обоих.

Два симплекса σ и $\bar{\sigma}$ являются k -связанными, если существует последовательность симплексов $\sigma, \sigma_1, \dots, \sigma_n, \bar{\sigma}$ такая, что любые два из них, следующие друг за другом, имеют общую k -грань, т.е. у них по меньшей мере

$(k+1)$ общая вершина. Такая цепь симплексов получила название k -цепи. Комплекс K является k -связанным, если любые два симплекса в K размерности больше или равной чем k являются k -связанными. Размерность комплекса равна максимальной размерности составляющих его симплексов.

Абстрактный комплекс K на конечном множестве $V = \{v_0, v_1, \dots, v_k\}$ вершин – это множество V с его непустым подмножеством булеана (или совокупностью его конечных подмножеств) S такое, что если $X \in S$ и $Y \subset X$, то $Y \in S$. При этом элементы множества V называются вершинами комплекса, а элементы множества S называются его симплексами σ . Другими словами, симплициальный комплекс K замкнут относительно взятия подмножеств: если $\sigma_T \in K$ и $S \subseteq T$, то $\sigma_S \in K$. Абстрактные симплициальные комплексы являются комбинаторными моделями определенных выше топологических симплициальных комплексов.

Построенный таким образом на основе исходных данных “каркас” удобно представлять в матричном виде, как показано на рис. 1, где столбцы соответствуют вершинам, а строки описывают симплексы комплекса. Тут множество вершин $V = \{1, 2, \dots, 12\}$, а сам симплициальный комплекс состоит из подмножеств $a = \{1, 2, 3, 4\}$, $b = \{3, 4, 5\}$, $c = \{5, 8\}$, $d = \{3, 6, 7\}$, $e = \{7, 8, 9, 10, 11\}$ и $f = \{9, 10, 11, 12\}$. Размерность комплекса равна четырем.

Рис. 1. Бинарное матричное представление комплекса, аппроксимирующего исходные данные посредством модели составленных в комплекс симплексов.

Пример (ячейка Вороного и триангуляция Делоне). Пусть $|yz|$ означает евклидово расстояние между двумя точками $y, z \in \mathbb{R}^d$. Ячейка Вороного для $p \in X, V_p = \{x \in \mathbb{R}^d : |px| \leq |qx|, q \in X\}$ – это множество точек, которые находятся от p не дальше, чем от какой-то другой точки q из нашего множества. Диаграммой Вороного $V_X = \{V_p : p \in X\}$ рассматриваемого множества является разбиение X на ячейки Вороного; k -мерная триангуляция Делоне – это абстрактный симплициальный комплекс с множеством вершин X , где семейство $\{x_0, x_1, \dots, x_k\}$ соединяется в k -симплекс тогда и только тогда, когда $V_0 \cap V_1 \cap \dots \cap V_k \neq \emptyset$ для всех $k \geq 0$. Геометрически это реализуется как k -мерная триангуляция Делоне выпуклой оболочки X .

Пример (комплекс Чеха). Построим симплициальный комплекс $\hat{C}_\varepsilon(X)$ с множеством вершин X так, что его подмножество $S \subseteq X$ будет являться симплексом, если шары $B_\varepsilon = B_\varepsilon(x, \varepsilon) = \{y \in \mathbb{R}^d : |xy| \leq \varepsilon, x \in S, \varepsilon \in \mathbb{R}\}$ с центрами в точках S имеют непустое пересечение: $B(x_0, \varepsilon) \cap B(x_1, \varepsilon) \cap \dots \cap B(x_k, \varepsilon) \neq \emptyset, \{x_0, x_1, \dots, x_k\} \in S$. Гомотопические типы $\hat{C}_\varepsilon(X)$ и объединения шаров радиуса ε с центрами в X совпадают [3, 4]. Тут мы имеем целое семейство симплициальных комплексов $\{C_\varepsilon : C_{\varepsilon_1} \subseteq C_{\varepsilon_2}, \varepsilon_1 \leq \varepsilon_2\}$, называемое фильтрованным комплексом (рис. 2).

Рис. 2. Фильтрация комплексов Чеха $C_\varepsilon(X)$ в $\mathbb{R}^d$, где индекс фильтрации ε является радиусом раздувающихся 2-мерных шаров B_ε .

Тут важно подчеркнуть, что именно на этапе конструирования, в зависимости от размера исходных ячеек, задается различная степень разрешения модели. Позже мы посмотрим, как ведут себя топологические особенности конструкции при движении по вложенной системе симплициальных комплексов с одного уровня разрешения на другой.

Разбиению множества и построениям разного рода конструкций на основе точек множества (комплекс Чеха, комплекс Виеториса-Рипса, триангуляция Делоне, альфа-комплекс и др.) посвящено множество статей, – например, [3, 5], разработаны соответствующие алгоритмы. Подобную приближа-

ющую конструкцию возможно построить не только на основе облака d -мерных чисел, но и аппроксимировать с различной степенью разрешения уже имеющиеся d -мерные объекты (рис. 3). Здесь надо упомянуть, что неориентированные связные поверхности в $\mathbb{R}$ не вкладываются, но разговор об ориентации многообразий выходит за рамки этой статьи.

Рис. 3. Различные уровни разрешения 3-мерной триангуляции Делоне объекта в $\mathbb{R}$.

Смысл алгебраических определений топологических понятий состоит в удобстве дальнейших вычислений. Так, абстрактный симплициальный комплекс легко моделируется явным цепным комплексом, k -й член которого порожден k -симплексами. Алгебраические, топологические и комбинаторные свойства симплициальных комплексов делают их универсальным средством для моделирования сложных структур, а также связности между различными подструктурами. На базе фильтрации абстрактных комплексов удобно вычислять интересующие нас и определяемые далее топологические характеристики. Ключевым моментом рассматриваемого в статье метода является установление этих топологических инвариантов с помощью приближенной математической модели. После предварительного моделирования данных становится возможным применение методов вычислительной топологии, при использовании которых сложность и разрешение модели варьируются в процессе ее построения. Конструкция представляет собой систему вложенных подмоделей, где вложения происходят как по уровню разрешения, так и по размерности. Путем вычисления наиболее стойких топологических инвариантов изменяющейся во времени конструкции находится их статистика, которая отражает информацию о связности исходных данных. Геометрический аналог для представления полученной информации служит наглядным фильтром для разделения истинных и преходящих топологических особенностей. Найденные топологические артефакты становятся объектом анализа и интерпретации.

За методами изначального восстановления и упорядочения входных данных, а также деталями реализации алгоритмов можно обратиться к работам [6, 7].

Математические модели и топологические методы

В последние десятилетия большое развитие получила вычислительная топология, в которой были разработаны методы вычисления так называемой персистентной, или устойчивой, гомологии, дающей возможность различать топологические особенности пространства при различных пространственных разрешениях [8, 9]. Каждый аппроксимационный комплекс строится постепенно из объектов возрастающей размерности, поэтому для каждого K фильтрация подкомплексов отражает наглядную пошаговую историю его конструирования. С течением времени топологические атрибуты разных размерностей возникают и исчезают в этой фильтрации, то удерживаясь в ней надолго, то почти мгновенно исчезая. Более стойкие топологические инварианты определяются посредством широкого спектра так называемой пространственной шкалы и служат для улавливания скорее истинных особенностей нижележащего пространства, чем артефактов облака данных и “шума”, зависящих от выбора тех или иных параметров разрешения. Пространственная шкала используется для описания или классификации при системной аппроксимации в зависимости от длины, расстояния, площади, а также прочих многомерных характеристик рассматриваемого опирающегося на эти данные объекта.

Для симплициального комплекса K в $\mathbb{R}$ определим k -цепь как подмножество k -симплексов из K следующим образом. Формально k -цепь – это линейная комбинация k -симплексов с целыми коэффициентами, которая может быть однозначно представлена как $c = c_k = \sum_{i=1}^{n_k} \alpha_i \sigma^i$, где n_k обозначает число k -симплексов в K , $\alpha_i \in \mathbb{Z}$; $\sigma^i = \sigma_k^i$ – i -й k -симплекс в комплексе. Определим сложение k -цепей c и d как коммутативную операцию симметрической разности двух множеств: $c + d = (c \cup d) - (c \cap d)$. Множество всех k -цепей вместе со сложением формирует группу C_k , чьим нулевым элементом является пустое множество. Объединение $(k-1)$ мерных граней k -симплекса $\sigma = \sigma_k$ является $(k-1)$ -мерной цепью, формирующей границу $\partial_k(\sigma) = \sigma \partial_k$ для σ , т. е. для $\sigma = \langle v_0, v_1, \dots, v_k \rangle \in K$ имеем $\sigma \partial_k = \sum_i (-1)^i \langle v_0, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k \rangle$, где v_i удален из последовательности. Тогда границей k -цепи будет сумма границ входящих в нее симплексов: $c \partial_k = \sum_{i=1}^{n_k} \alpha_i (\sigma^i \partial_k)$. Граница k -цепи сама границы не имеет: $\partial_{k-1}(c \partial_k) = 0$. Каждый граничный оператор является гомоморфизмом

$\partial_k : C_k \rightarrow C_{k-1}$ между группами k -й и $(k-1)$ -й цепей, а при $k = 0, 1, \dots, d$ совокупность граничных операторов соединяет цепные группы в цепной комплекс $\emptyset \rightarrow C_d \xrightarrow{\partial_d} C_{d-1} \xrightarrow{\partial_{d-1}} C_{d-2} \xrightarrow{\partial_{d-2}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} \emptyset$, где $\partial_{k-1} \partial_k = 0$.

Два объекта, определяемые через k -ю циклическую и k -ю граничную группы соответственно, представляют тут особый интерес:

1) k -циклы $Z_k = \text{Ker}(\partial_k) = \{z \in C_k : z \partial_k = 0\}$ – набор k -цепей с пустой границей;

2) k -границы $B_k = \text{Im}(\partial_{k+1}) = \{b \in C_k : b = d \partial_{k+1}, d \in C_{k+1}\}$ – собрание k -цепей, являющихся границей для $(k+1)$ -цепей.

Группы Z_k и B_k обе являются k -цепями в $\text{Ker}(\partial_k)$ и $\text{Im}(\partial_{k+1})$ соответственно, вместе с операцией сложения они формируют подгруппы C_k , а из $\partial_k \partial_{k+1} = 0$ следует, что эти группы вкладываются друг в друга: $B_k \subseteq Z_k \subseteq C_k$. Граница любой вершины есть пустое множество, поэтому каждая 0-цепь является 0-циклом: $C_0 = Z_0$.

Каждый элемент из $\text{Ker}(\partial_k)$ является подкомплексом без границы, так что он представляет собой k -мерный цикл. В то время как элементы из $\text{Im}(\partial_{k+1})$ являются k -границами для цепей большей размерности, k -циклы являются для нас ключевыми топологическими объектами, определяющими устойчивость k -мерных дыр симплициального комплекса; k -я группа гомологий определяется как $H_k = Z_k / B_k = \text{Ker}(\partial_k) / \text{Im}(\partial_{k+1})$, ее элементами являются классы гомологий $z + B_k = \{z + b : b \in B_k\}$ для всех $z \in Z_k$, т. е. два цикла $z_1, z_2 \in Z_k$ гомологичны, $z_1 \sim z_2$, если $z_1 - z_2 \in B_k$. Нулевым элементом группы H_k является $\emptyset + B_k = B_k$, а сумма двух классов есть $(z_1 + B_k) + (z_2 + B_k) = (z_1 + z_2) + B_k$.

При этом k -е числа Бетти симплициального комплекса K определяются как ранги k -х групп гомологий: $\beta_k = \beta(H_k) = \text{rank}(H_k) = \text{rank}(Z_k) - \text{rank}(B_k)$. Для комплекса из $\mathbb{R}$ ненулевыми β_k могут быть только при $0 \leq k \leq d-1$. Конструирование гомологий подразумевает, что мы избавляемся от циклов, являющихся границами подкомплексов большей размерности. Поэтому оставшиеся несут в себе информацию о k -мерных дырах комплекса, а β_k равны числу этих дыр. Используя двойственность Александера [10], можно проиллюстрировать первые числа Бетти следующим образом. Не являющиеся границами 0-циклы представляют множество компонентов комплекса, один базисный элемент на компонент, поэтому β_0 равно числу компонентов K ; 1-циклы, которые сами не границы, представляют комплект нестягиваемых замкнутых

кривых в K , поэтому β_1 равно числу формируемых K туннелей; 2-циклы, не являющиеся границами, подразумевают множество нестягиваемых замкнутых поверхностей в K , поэтому β_2 равно числу внутренних пустот в K , которые есть элементы дополнения K , т.е. элементы $\mathbb{R} - K$ и т.д. Поскольку K из $\mathbb{R}$, то не существует непустых d -циклов или d -границ, т.е. $Z_d = B_d = \{\emptyset\}$.

Симплициальный комплекс конструируется, начиная с пустого множества, путем пошагового добавления симплексов. Возникающая при этом последовательность вложенных подкомплексов называется фильтрацией. Для каждой актуальной размерности k фильтрацией симплициального комплекса K_k является упорядоченная последовательность комплексов K_k^i , скрепленных отображениями вложения $K_k^i \rightarrow K_k^{i+1}$, т.е. имеем $\emptyset = K_k^0 \subset K_k^1 \subset \dots \subset K_k^n = K_k$, где верхний индекс фильтрации может интерпретироваться как время. Заметим, что возможно не только добавлять симплексы в комплекс в процессе построения фильтрации, но и удалять их. Для измерения продолжительности жизни определенных топологических особенностей комплекса в течение истории его конструирования нам придется расширить понятие вышеопределенных гомологий.

Помимо прочего, на K_k^i действуют граничные операторы $\partial_i : K_k^i \rightarrow K_{k-1}^i$, где понижение размерности комплекса означает понижение максимальной размерности образующих его симплексов. Рассмотрим расширенную фильтрацию симплициальных комплексов, в которой отражаются также переходы между размерностями комплексов. Ее удобно изображать в виде диаграммы, где индекс фильтрации возрастает слева направо, а размерности убывают сверху вниз по мере действия граничных операторов (рис. 4).

$$\begin{array}{ccc}
 \vdots & \vdots & \vdots \\
 \partial_3 \downarrow & \partial_3 \downarrow & \partial_3 \downarrow \\
 \emptyset = K_2^0 \subset K_2^1 \subset K_2^2 \subset \dots & & \\
 \partial_2 \downarrow & \partial_2 \downarrow & \partial_2 \downarrow \\
 \emptyset = K_1^0 \subset K_1^1 \subset K_1^2 \subset \dots & & \\
 \partial_1 \downarrow & \partial_1 \downarrow & \partial_1 \downarrow \\
 \emptyset = K_0^0 \subset K_0^1 \subset K_0^2 \subset \dots & &
 \end{array}$$

Рис. 4. Расширенная фильтрация комплексов по уровню разрешения и по размерности.

Нас интересуют топологические особенности, сохраняющиеся неизменными в последовательности симплициальных комплексов различных габаритов: мы смотрим, как при движении по фильтрации ведут себя ранги

групп гомологий, являющиеся хоть и не исчерпывающими, но все же важными неотъемлемыми характеристиками нашей модели. Гомологии выявляют не являющиеся границами циклы, которые важны для нас только в том случае, если время их пребывания в комплексе превышает некоторый заранее заданный порог p . Другими словами, эти циклы все еще должны присутствовать в комплексе, находящемся через p шагов далее в фильтрации, после момента их возникновения. Циклы, конвертирующиеся в границы быстрее установленного временного порога, игнорируются. Таким образом, определяется устойчивость, а значит, и истинность топологических особенностей комплекса. И тут необходим алгебраический “сепаратор”, который позволил бы выделить эти атрибуты с относительно продолжительной жизнью в истории развертывания фильтрации комплексов. Скомбинируем группы гомологий всех этих комплексов в единую алгебраическую структуру следующим образом. Применение k -мерного гомологического функтора H_k к объектам и отображениям расширенной фильтрации даст соответствующую систему групп гомологий, т.е., множество абелевых групп $H_k(K^i) = H_k^i$ вместе с гомоморфизмами $\lambda_k^i = \lambda_k^{i,1} : H_k^i \rightarrow H_k^{i+1}$, вкладывающими один гомологический класс в другой, его содержащий. В построенной модели $H_k(K)$ обозначает число k -мерных подкомплексов K , которые не имеют границы в K и сами не являются границей ни для каких $(k+1)$ -мерных подкомплексов. Для каждой актуальной размерности фильтрации симплициальных комплексов сопоставляется соединенная гомоморфизмами последовательность групп гомологий $0 = H_k^0 \xrightarrow{\lambda_k^0} H_k^1 \xrightarrow{\lambda_k^1} \dots \xrightarrow{\lambda_k^{n-1}} H_k^n \xrightarrow{\lambda_k^n} H_k^{n+1} = 0$, где добавление нулевой группы в конце означает, что любой гомологический класс в итоге исчезает.

Напомним, что функтор $F: I \rightarrow J$ из категории I в категорию J – это отображение, которое:

- 1) сопоставляет каждому объекту $U \in I$ объект $F(U) \in J$;
- 2) сопоставляет каждому морфизму $f: U \rightarrow V$ в категории I морфизм $F(f): F(U) \rightarrow F(V)$ в категории J .

Причем это сопоставление таково, что:

- а) $F(id_A) = id_{F(A)}$;
- б) $F(g \circ f) = F(g) \circ F(f)$.

Тождественные морфизмы и структуру композиции морфизмов функтор сохраняет.

Пусть теперь в l -м комплексе фильтрации Z_k^l и B_k^l представляют k -ю циклическую и граничную группы соответственно. Для вычисления “долго-

играющих” и при этом не являющихся границами циклов мы рассмотрим фактор Z_k^l по граничной группе комплекса K_k^{l+p} , находящегося на p шагов дальше в фильтрации после K_k^l . Другими словами, нас интересует вычисление гомологий не для одного, а сразу для желаемо большого количества подкомплексов в фильтрации.

Рассмотрим последовательность из p элементов фильтрации, скрепленную вместе инъекциями $K_k^l \rightarrow K_k^{l+p}$. После применения функтора H_k получаем гомоморфизм $\lambda_k^{l,p} : H_k^l \rightarrow H_k^{l+p}$. Персистентная гомология определяется как образ этого гомоморфизма, а алгебраически выражается как $H_k^{l,p} = Z_k^l / (B_k^{l+p} \cap Z_k^l)$, т.е. p -стойкая k -я группа гомологий l -го комплекса в фильтрации определяется как фактор-группа его k -й циклической группы по отношению к k -й граничной группе, посчитанной в фильтрации на p комплексов позже. Определение корректно, потому что $B_k^{l+p} \cap Z_k^l$, будучи пересечением двух подгрупп C_k^{l+p} , само является группой и к тому же подгруппой Z_k^l .

Поясним, что это означает. Мы получаем циклы, которые появились в K_k^l или раньше и не обращаются в границы минимум в течение p единиц времени (или шагов в фильтрации), начиная с момента l , т.е. эти циклы все еще присутствуют и независимы в K_k^{l+p} . Если какие-то два цикла гомологичны в K^l , то они также существуют и гомологичны в K^{l+p} , а следовательно, мы имеем гомоморфизм $\lambda_k^{l,p} \cong H_k^{l,p}$. Определим теперь p -стойкие k -е числа Бетти l -го комплекса K_k^l в фильтрации как $\beta_k^{l,p} = \text{rank}(H_k^{l,p})$. Они равны числу гомологических классов, возникших в фильтрации в момент времени не позже чем l и остающихся там до момента времени не раньше чем $l + p$. Более глобально, $\beta_k^{l,p}$ отражают регулируемую информацию об устойчивой k -мерной связности K .

Интерпретации

Над полем для каждой размерности k группы гомологий $H_k(K)$ комплекса K являются вектор-пространствами, с точностью до изоморфизма описываемые своими рангами. Персистентная гомология является алгебраическим инвариантом, идентифицирующим рождение и смерть каждого релевантного топологического атрибута в сконструированной модели. Топологические особенности фильтрации комплексов имеют свои продолжительности жизни внутри этой аморфной конструкции, в течение которого они отображаются посредством соответствующих чисел Бетти. Эти отражения вспыхивают и гаснут в процессе развертывания процесса во времени: k -е числа Бетти озна-

чают, что где-то в построенной системе фильтрации комплексов k -размерностный подкомплекс выпадает сквозь все стадии роста или редукции в течение времени p . Поэтому некий объект, сформированный симплексами размерности не выше чем k , отсутствует в комплексе указанное время. Такие k -размерностные дыры остаются открытыми, пока мы утолщаем комплекс в фильтрационной цепи как минимум до индекса фильтрации p .

Интерес представляют атрибуты с наибольшей продолжительностью жизни, поскольку они дольше других выживают в построенной модели, то они и являются истинными особенностями изначального объекта рассмотрения. Очень наглядно для каждой размерности представлять эти временные промежутки жизни соответствующих топологических особенностей в графическом виде: набора интервалов $\{[a_l, b_l) : a_l \in \mathbb{Z}_+, b_l \in \mathbb{Z}_+ \cup \{\infty\}\}$, получивших название p -интервалов, группируемых в штрих-коды. Тут каждому классу гомологий графически соответствует отрезок штрих-кода, где концы отрезка обозначают время появления и исчезновения класса. Начиная со времени l , число таких отрезков длины не короче чем p для H_k равняется $\beta_k^{l,p}$, где параметр p целиком определяется выбранным уровнем разрешения. Мера устойчивости генераторов k -й группы гомологий графически соответствует длинам соответствующих p -интервалов. Мы получили своего рода математический эквалайзер для выправления топологического “сигнала”. Изменяя параметр p , можно регулировать объем отсекаемого топологического “шума”, которым считаются топологические атрибуты с временем жизни, не превышающим выбранного порога.

Так как числа $\beta_k^{l,p}$ описывают многомерную связность модели, то посредством этого также дается описание связности изначального облака чисел. Это соответствие гарантируется принципами построения аппроксимационной конструкции на основе исходных данных. В итоге мы получаем возможность разглядеть структуру данных, содержащую в себе важную информацию о представленном этими данными объекте или процессе. Алгоритм вычисления устойчивых чисел Бетти реализован автором на языке SINGULAR [11] в [7]. Использование вышеописанного метода исследования дает возможность анализировать и интерпретировать числовые массивы данных, что и будет проиллюстрировано на конкретных примерах.

Примеры

Допустим, мы хотим разработать лекарство от определенной болезни для воображаемого “среднего” пациента госпиталя. Нам нужна математическая модель, в которой каждый пациент был бы закодирован как точка в мно-

гомерном пространстве с координатами – всеми возможными сопутствующими характеристиками рассматриваемого пациента. Например: пол $\rightarrow \{0,1\}$, возрастная категория по десятилетиям $\rightarrow \{1-9\}$, статистика обращений и время, проведенное в стационаре $\rightarrow \{numbers\}$, перенесенные болезни $\rightarrow \{code\}$, число операций $\rightarrow \{number\}$, наследственные заболевания $\rightarrow \{code\}$, проведенные курсы лечения $\rightarrow \{code\}$ и прочее. Далее мы смотрим, где соответствующий “средний” пациент, на которого мы ориентируемся, расположен относительно наших реальных входных данных по пациентам (рис. 5), где сверху: множество данных; внизу: симплициальный комплекс на основе данных. Дыра в данных слева.

Рис. 5. Иллюстрация использования информации о связности оцифрованных исходных данных в $\mathbb{R}$

Если вышло так, что он попадает в дыру в данных по всем или по большинству параметров, то выбранный нами тестовый пациент с соответствующим набором характеристик находится скорее в стороне от реальных пациентов, и тогда разрабатываемое лекарство едва ли будет эффективным. Поэтому бесполезно ориентироваться на данного абстрактного пациента, а стоит сконцентрироваться на каком-то более близком к реальности, которого имело бы смысл назначить “средним”, т.е. нам нужна точка в многомерном пространстве, расположенная близко если не ко всем сразу, то хотя бы ко многим точкам из реальных входных данных. Таким образом мы избегаем провалов в дыры исследуемого рыхлого облака многомерных чисел. Следовательно, разрабатываемое лекарство, весьма вероятно, возымеет эффект хотя бы на некоторую группу реальных пациентов, располагающихся в рамках по-

строенной модели, достаточно близко к выбранной точке ориентации.

Подобные рассуждения годятся также при разработке стратегии рекламы, ориентированной на “среднего” потребителя, или при составлении плана предвыборной агитации в избирательной кампании, направленной на некоторого среднестатистического избирателя. Не существует универсальной методики, которая бы годилась для всех сразу, но возможна техника воздействия на узконаправленные группы клиентов. Как было показано, многомерные дыры в данных имеют важное значение, потому что отражают информацию о структуре данных. Следовательно, вычисления многомерной связности, присущей построенной на основе исходного облака чисел модели, дают возможность предсказаний об эффективности той или иной методики воздействия или поведения по отношению к представляемому этими данными объекту или процессу.

Рассмотрим теперь пример использования многомерного анализа при оценке ликвидности акций компаний. Допустим, мы хотим оценить целесообразность и риски при покупке рыночных акций. Информация, которая помогла бы сделать такую оценку, должна быть переведена в цифровой формат. Параметрами для такого перевода снова могут послужить некоторые базисные характеристики: число работников компании, ценность различных активов, стоимость сырья, процент выгоды, размер таможенных пошлин при импорте/экспорте, капитализация и т.д. Все эти числовые данные, взятые в последовательности, могут быть интерпретированы как точки в соответствующем многомерном пространстве. Сравнительный анализ множества других подобных компаний по всему миру даст возможность выделить регион описанного пространства, в котором покупка акций сопряжена с наименьшими рисками.

Заключение

Топологическая модель данных представляет собой систему вложенных симплициальных комплексов, где вложения происходят как по уровню разрешения, так и по размерности. Построенные на базе такой фильтрации комплексов персистентные гомологии были определены как гомологии расширяющегося пространства, отражающие время жизни присущих объекту рассмотрения топологических атрибутов. Особенное удобство заключается в кодировании устойчивых групп гомологий в форме параметризованной версии их рангов. Время жизни генераторов этих групп в фильтрации представляется геометрически как множество интервалов. Штрих-коды параметризуют устойчивость чисел Бетти и служат фильтрами, позволяющими различать

присущую комплексу истинную и шумовую информацию, т.е. дают возможность выявить в построенной упрощенной модели дыры разного порядка. Согласно принципам моделирования, многомерная связность фильтрации комплексов и облака входных данных идентичны.

Использование упрощенных моделей для анализа данных является мощным инструментом для глубокого проникновения в природу представленных этими данными объектов, процессов или явлений. Как мы могли убедиться, даже простые аппроксимационные симплексные конструкции позволяют пролить свет на основные закономерности в рассматриваемом числовом облаке данных, а также на взаимосвязи между подкластерами таких данных. В свою очередь это ведет к возможности более эффективных подходов в использовании представленных систем, а также к способности делать предсказания насчет будущего рассматриваемых процессов.

Отметим, что даже неполные и/или испорченные “шумом” данные все еще содержат массу полезной информации, которая может быть просчитана после предварительного восстановления и/или выправления данных с приемлемым уровнем аппроксимации. Стоит иметь в виду, что обработка гигантских массивов реальных данных требует значительного времени, поэтому использование мощных компьютеров весьма желательно в этом случае. Более сложное моделирование повышает количество и улучшает качество возможных интерпретаций. Но привлечение таких сложных методов требует аккуратности в конструировании и способности истолковывать полученные результаты в порядке обратного перевода с языка построенной модели на язык жизненных реалий.

ЛИТЕРАТУРА

1. *Edelsbrunner H.* The union of balls and its dual shape // *Discrete and Computational Geometry*. – 1995. – Vol. 13. – P. 415-440.
2. *Leray J.* Sur la forme des espaces topologiques et sur les points fixes des représentations // *J. Math. Pures Appl.* – 1945. – Vol. 24(9). – P. 95-167.
3. *Edelsbrunner H., Shach N.R.* Triangulating topological spaces // *International Journal of Computational Geometry & Applications*. – 1997. – Vol. 7(4). – P. 365-378.
4. *Kim J., Shin J., Chazal F., Rinaldo A., Wasserman L.* Homotopy reconstruction via the Čech complex and the Rips complex // *arXiv.org*. – 30 December 2019. URL: <https://arxiv.org/abs/1903.06955>.
5. *Carlsson G., Silva V.* Topological approximation by small simplicial complexes // *Semantic Scholar*. — 2003. URL: <https://www.semanticscholar.org/paper/Topological-approximation-by-small-simplicial-Carlsson-Silva/1f41d1e30ac2dc791c6b6b085e70bf9fee64d5ce>.
6. *Artamonov O.* Predictions based on the rectification and processing of noisy data // *IEEE Computer Society Digital Library (CSDL) and I-Xplore*. – 02 August 2018. URL:

<https://ieeexplore.ieee.org/document/8424306>.

7. *Artamonov O.* Topological methods for the representation and analysis of exploration data in oil industry // University of Kaiserslautern. – 30 August 2010. URL: <https://kluedo.ub.uni-kl.de/frontdoor/index/index/docId/2234>.
8. *Zomorodian A., Carlsson G.* Computing persistent homology // Discrete and Computational Geometry. – 2005. – Vol. 33(2). – P. 249-274.
9. *Carlsson G., Zomorodian A.* The theory of multidimensional persistence // Discrete and Computational Geometry. – 2009. – Vol. 42. – P. 71-93.
10. *Hatcher A.* Algebraic topology. – Cambridge University Press, 2002.
11. SINGULAR: система компьютерной алгебры [Свободное программное обеспечение, созданное в рамках проекта GNU]. – Режим доступа: <https://www.singular.uni-kl.de>

Статья предоставлена к публикации членом редколлегии Е.А. Нурминским.

E-mail:

Артамонов Олег Николаевич – artamonov.on@dvfu.ru, grtmnv@gmail.com.