Алгоритм проверки статистической значимости кластерной структуры последовательностей на примере анализа гликемических временных рядов

Кладов Д.Е., Бериков В.Б., Климонтов В.В.

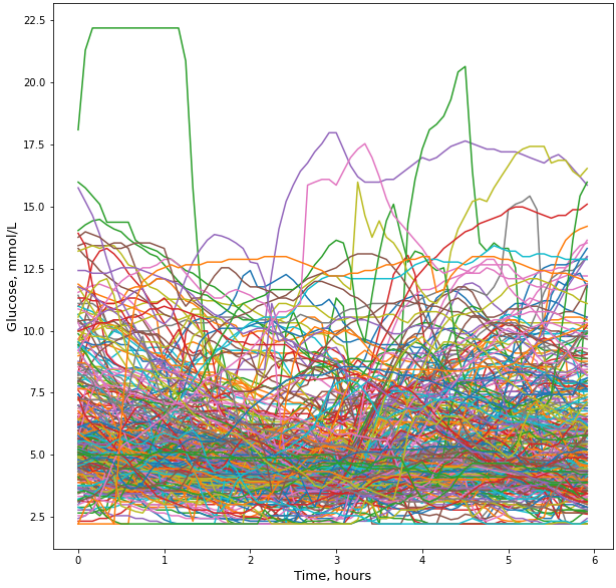
Новосибирский государственный университет
Научно-исследовательский институт клинической и экспериментальной лимфологии
Институт математики им. С. Л. Соболева СО РАН

Научная конференция "Современные проблемы обратных задач", посвященная 90-летию со дня рождения академика М.М. Лаврентьева. Новосибирск, Академгородок, 19 - 23 декабря 2022 года

Постановка задачи и цели

- Пусть задано некоторое множество объектов, которые являются временными рядами;
- Требуется разработать алгоритм кластеризации временных рядов;
- Требуется разработать метод оценивания статистической достоверности кластеризации;
- Требуется исследовать метод на реальных гликемических кривых и оценить качество кластерной структуры, полученной в результате применения алгоритма иерархической кластеризации.

Гипогликемические временные ряды



Задача кластеризации временных рядов

Требуется разбить множество временных рядов на группы схожих между собой объектов путём максимизации (минимизации) целевого критерия:

$$N = N_1 \cup \cdots \cup N_s,\ N_i \cap N_j = \varnothing,\ i \neq j$$

$$F\left(N_1, \ldots, N_s\right) \to max$$

N – множество объектов из обучающей выборки, N_i – объекты из кластера i, s – количество кластеров.

Выбор метрик для кластеризации

- Между объектами
  - **Euclidean**
  - Manhattan
  - Hamming
- Между кластерами
  - **Ward method**
  - Single
  - Complete
  - Weighted
  - Centroid
  - Median
- Внутри кластеров
  - Diametric
  - **Variance**
  - Centroid

Кластеризация временных рядов

- Выделение признаков
  - Discrete Fourier Transform (DFT)
  - Discrete Wavelet Transform (DWT)
- Попарное сходство
  - Dynamic Time Warping (DTW)
  - Longest Common SubSequence (LCSS)
- И другие

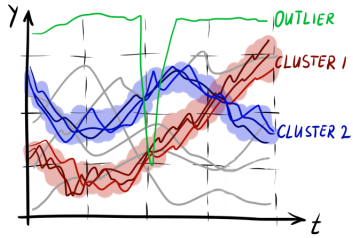


Рис 1. Пример кластеризации временных рядов

Методы оценивания качества кластеризации

- Индекс качества silhouette score

Для объекта i из кластера C_I определим среднее расстояние до объектов из кластера C_I : $a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i,j)$ и среднее расстояния до объектов из ближайшего кластера C_J : $b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$. Тогда индекс силуэт для объекта i определяется следующим образом: $s(i) = \frac{b(i)-a(i)}{max(b(i),a(i))}$; для всего набора данных: $S = \frac{1}{N} \sum_{i=1}^{N} s(i)$, $N-$объем выборки.

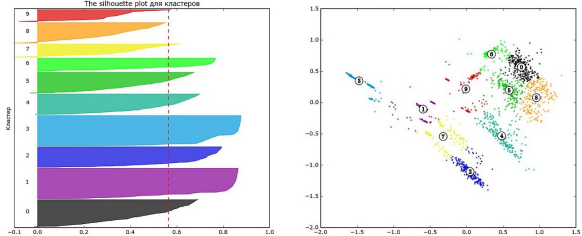


Рис 2. Пример индекса качества Silhouette score для объектов выборки

Методы оценивания качества кластеризации

- Метод Монте-Карло для проверки статистических гипотез

Выдвигается основная гипотеза о том, что в данных нет кластерной структуры, или, что то же самое, о существовании одного единственного кластера. Способ оценивания статистической достоверности полученной кластеризации заключается в сравнении качества кластеризации на реальной выборке с качеством кластеризаций на искусственно сгенерированных выборках временных рядов с теми же самыми числом объектов и их длиной.

Задача проверки статистической значимости кластеризации временных рядов

- X_0 – исходное множество временных рядов, $X_1, \ldots, X_t$ – t наборов множеств временных рядов, сгенерированных при условии выполнения основной гипотезы;
- требуемое количество кластеров k;
- количественная мера того, насколько разбиение на кластеры "хорошее", например, некий индекс качества: $S(X_i, k) = s_{i,k}$ – значение индекса на множестве временных рядов X_i при разбиении его на k кластеров;
- уровень значимости α.

Имея 1– 4, требуется оценить статистическую достоверность полученной кластерной структуры. Если $s_{0,k} > q_\alpha(s_{1,k}, \ldots, s_{t,k})$, где $q_\alpha(X)$ – квантиль уровня α множества X, то кластеризация принимается достоверной, в противном случае – нет.

Метод проверки статистической значимости кластеризации[1]

Идея метода: сгенерировать псевдовыборки, соответствующие основной гипотезе, и многократно провести кластеризацию на искусственно сгенерированной выборке. Затем определить, для какого количества кластеров индекс качества на реальных данных окажется больше квантиля уровня p для искусственных данных. В этом случае кластеризация принимается достоверной.

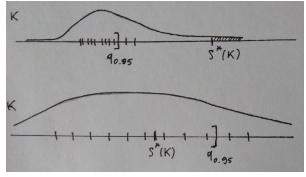


Рис 3. Метод Монте-Карло для проверки статистических гипотез

[1] I. Kirilyuk, O.Senko.Assessing the validity of clustering of panel data by Monte Carlo methods // COMPUTER RESEARCH AND MODELING. 2020 VOL. 12 №6 P. 1501–1513.

Метод проверки статистической значимости кластеризации

Было показано[2], что у всех временных рядов из набора данных есть значимая автоковариация для первых 3-х лагов. *Идея улучшения*: при генерации временных рядов учитывать автоковариацию, усреднённую по всему набору данных. Генерация временных рядов производилась из многомерного нормального распределения $N(E,C)$ где E-вектор выборочных средних, C-выборочная ковариационная матрица.

[2]Klimontov V., Kladov D., Berikov V., Semenova J. Nocturnal glucose fluctuations in patients with type 1 diabetes: which patterns are associated with hypoglycemia? // Diabetes Technology and Therapeutics. 2022.2525.abstracts

Алгоритм оценивания статистической достоверности кластеризации

Вход: целочисленный набор данных, α - уровень значимости, s - количество кластеров, метод вычисления индекса качества кластеризации, t - количество повторений в методе Монте-Карло

Выход: true, если кластеризация достоверна, false - иначе.

Шаги:

1. определить значение индекса качества кластеризации для реальных данных;
2. сгенерировать k временных рядов длины n согласно основной гипотезе;
3. определить значение индекса качества кластеризации для искусственных данных для s кластеров;
4. повторить шаги 2-3 t раз;
5. вычислить квантиль уровня $1 - \alpha$ для набора индексов качества искусственных данных.

Если значение индекса качества кластеризации для реальных данных окажется больше квантиля уровня $1 - \alpha$ для искусственных данных, то вернуть true, иначе вернуть false.

Набор данных

В качестве выборки был использован реальный набор данных, предоставленный НИИКЭЛ СО РАН, состоящий из временных рядов, полученных в результате непрерывного мониторинга уровня глюкозы у 385 взрослых пациентов с сахарным диабетом 1 типа. Временной ряд каждого пациента был разделён на ночные, ранние утренние и дневные периоды.

Таблица 1. Объёмы обучающих выборок

	Ночные вр.ряды	Ранние утренние вр. ряды	Дневные вр. ряды
N_{init}	2956	3036	2954
N_{final}	2782	2729	2156
$\Delta, \%$	5.9	10.1	27
N_{gl}	2464	2544	1630
N_{hg}	318	185	526

Результаты

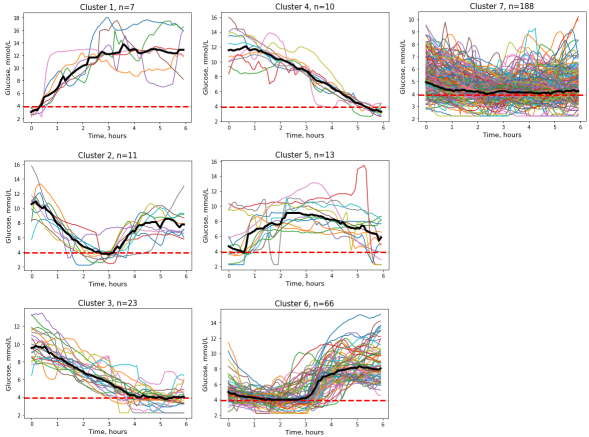


Рис 4. Кластеризация временных рядов с гипогликемией. Черная линия соответствует медоиду кластера, красная пунктирная линия соответствует порогу гипогликемии (3.9 ммоль/л)

Результаты

Метод оценки статистической достоверности полученных решений доказал надежную кластеризацию на уровне значимости $\alpha < 0{,}05$. Параметры алгоритма: количество временных рядов - 318, размерность - 72, количество повторений - 5000.

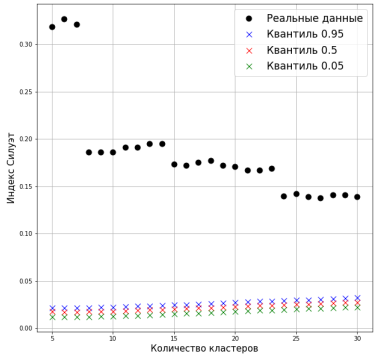


Рис 5. Зависимость индекса силуэт от числа выделенных кластеров.

Выводы

- Разработан метод проверки статистической значимости решений для произвольного алгоритма кластеризации;
- Разработана программная реализация метода;
- Разработанный метод был протестирован на реальных данных, предоставленных НИИКЭЛ СО РАН;
- Проведена кластеризация гликемических кривых
- С помощью метода доказана статистическая достоверность кластеризации;
- Определены паттерны флуктуаций уровня глюкозы у больных сахарным диабетом 1 типа в различные периоды суток.

Исследование выполнено за счёт гранта Российского научного фонда 20-15-00057.

Спасибо за внимание!