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, Новосибирск
Научно-исследовательский институт клинической и экспериментальной лимфологии – филиал ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск, Россия
ФГБУН институт математики им. С. Л. Соболева Сибирского отделения Российской академии наук
danil_kladov@mail.ru

В данной работе рассматривается задача кластеризации временных рядов с помощью алгоритма иерархического кластерного анализа, а также задача проверки статистической значимости полученной кластерной структуры. Для определения статистической достоверности проверяется основная гипотеза о том, что в данных нет кластерной структуры. Проверка осуществляется с помощью метода Монте-Карло, который основан на многократной кластеризации искусственно сгенерированной выборки из заданного распределения, вычислении индекса качества кластерной структуры и сравнении с индексом качества кластеризации на реальных данных. Кластеризация принималась достоверной для данного числа выделенных кластеров, если значение индекса на реальной выборке оказывалось больше значения 95%-ного квантиля для искусственных данных.

В качестве выборки реальных данных использован набор данных, состоящий из временных рядов, которые были получены в результате непрерывного мониторинга глюкозы у 385 взрослых пациентов с сахарным диабетом 1 типа. Измерения проводились с интервалом времени 5 минут. Данные предоставлены ИЦиГ СО РАН. Гипогликемическими считались те временные ряды, в которых уровень глюкозы достигал значения < 3.9 ммоль/л хотя бы 15 минут подряд. Для анализа использовались ночные, ранние утренние и дневные промежутки времени: 0.00-5.59, 4.00-7.59 и 8.00-23.59 соответственно. На этих данных предложенная методика показывает достоверную кластеризацию на уровне значимости $p < 0.05$.

По результатам кластеризации в ночные, ранние утренние и дневные промежутки времени получилось 15, 11, 7 кластеров без гипогликемии и 7, 4, 5 кластеров с гипогликемией соответственно. Различия в кластерах касались начального и конечного уровня глюкозы, наличия или отсутствия нисходящего/восходящего тренда. Прослеживалась зависимость между исходным значением глюкозы, быстрой падения и временем возникновения гипогликемии.

Данное исследование поддержано грантом РНФ (20-15-00057).