Вычислительная реализация рангового метода кластеризации

Гренкин Глеб Владимирович
Дальневосточный Федеральный Университет (Владивосток), Россия
e-mail: glebgrenkin@gmail.com

Кластеризация — это разбиение совокупности объектов на классы, в каждом из которых объекты похожи друг на друга. При кластеризации объектов обычно выбирается метрика, определяющая меру сходства объектов.

В статье [1] был предложен ранговый метод кластеризации, в котором не используется метрика. В статье [1] авторы рассматривают одномерные эмпирические данные: $w_1, w_2, \ldots, w_n$, — эти данные упорядочиваются по возрастанию, и каждому значению w ставится в соответствие порядковый номер — ранг r. Исходные данные анализируются с помощью соотношения

$$\ln w \cong -\gamma \ln\left(\frac{N-r}{r}\right) + c \equiv -\gamma \ln R + c, \tag{1}$$

которое представляет собой модифицированный В. П. Масловым закон Ципфа (в [1] принято $N = 2n + 1$). При разбиении данных на кластеры ранговым методом на каждом из кластеров справедлив модифицированный В. П. Масловым закон Ципфа со своими значениями параметров (γ, c), которые меняются при переходе от кластера к кластеру.

Требуется автоматизировать процесс кластеризации эмпирических данных ранговым методом.

Разработана математическая модель, дающая ответ на вопрос: какую информацию должна выводить программа? Предлагается найти так называемое множество максимальных промежутков, которое позволяет разбить данные на кластеры. При этом выбор конкретного разбиения остаётся за пользователем (здесь пользователь может использовать априорную информацию).

Возникает задача нахождения минимальной высоты полосы, содержащей все точки некоторого промежутка. При этом могут присутствовать аномальные точки — точки, которые можно выбросить. Для решения этой задачи получено обобщение алгоритма Грэхема построения выпуклой оболочки [2].

Разработана программная система [3], в которой реализованы разработанные методы.

Список литературы

[1] Гузев М. А., Черныш Е. В. Ранговый анализ в задачах кластеризации // Информатика и системы управления. 2009. №3(21). С. 13–19.

[2] Гренкин Г. В. Методы вычислительной реализации рангового метода кластеризации // Информатика и системы управления. 2012. № 1(31). С. 71–79.

[3] Гренкин Г. В., Черныш Е. В. ClRank // Свидетельство Роспатента о государственной регистрации программы для ЭВМ. — Рег. № 2012612452 от 06.03.2012.