Секция «Дискретная математика и математическая кибернетика»

Алгоритм кластеризации, основанный на интервальных формальных понятиях и приближенном решении задачи о поиске оптимального положения бруса

Научный руководитель – Галатенко Алексей Владимирович

Нерсисян С.А.[1], *Панкратьева В.В.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Механико-математический факультет, Москва, Россия; 2 - Московский государственный университет имени М.В.Ломоносова, Механико-математический факультет, Кафедра математической теории интеллектуальных систем, Москва, Россия

В настоящей работе рассматривается геометрический подход к задаче кластеризации, которая имеет важные приложения в машинном обучении, биоинформатике, обработке изображений и других областях. Под кластеризацией понимается задача представления множества $X \subset \mathbb{R}^d$ в виде такого дизъюнктного объединения подмножеств (кластеров), чтобы элементы из одного кластера были схожи (по какому-то критерию), а элементы из разных кластеров различались (см., например, [1]).

Один из известных методов кластеризации опирается на использование интервальных формальных понятий. Интервальные формальные понятия являются частным случаем узорных структур, введенных в работе [4]. Этот метод позволяет кластеризовать строки числового формального контекста (матрицы данных). При этом критерий близости строк заключается в том, что значения на соответствующих позициях попадают в наперед заданный интервал.

Данная проблема аналогична известной задаче из области вычислительной геометрии: требуется по заданному n-элементному набору точек в пространстве $\mathbb{R}^d$ найти такое положение бруса с фиксированными длинами сторон, чтобы он покрывал максимальное количество точек из набора (это положение называется оптимальным). Под брусом понимается декартово произведение одномерных отрезков, длины которых и называются длинами сторон бруса.

Задача поиска бруса и её вариации (например, поиск оптимального положения сферы) хорошо изучены в плоском случае ($d = 2$), для которого построены нижние и оптимальные верхние оценки (см. [3]). Тем не менее, в высоких размерностях не известны ни нижние оценки, ни алгоритмы, способные точно решить задачу за разумное время.

В настоящей работе предложен алгоритм, приближенно решающий задачу поиска оптимального бруса с временной сложностью

$$O(dn\log(n) + \frac{d^3 n^{1-\frac{1}{d}}}{s_{\min}} f(n,d))$$

(в худшем случае) и пространственной сложностью $O(n)$. Число s_{min} и функция f являются параметрами алгоритма: f обозначает количество итераций, которое необходимо выполнить в основном блоке алгоритма, а s_{min} регулирует их продолжительность. При увеличении количества и продолжительности итераций возрастает точность алгоритма. В качестве функции можно взять, к примеру, $f(n,d) = [\log(dn)]$ ($[x]$ — целая часть числа x).

Тестирование алгоритма кластеризации, основанного на вышеописанном приближенном решении задачи, на наборе данных The Cancer Cell Line Encyclopedia (см. [2]) показало хороший результат: полученное разбиение согласуется с априори известными биологическими факторами.

Источники и литература

1) Мандель И.Д. Кластерный анализ. М.: Финансы и статистика. 1988.
2) Barretina J., Caponigro G., Stransky N. et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity // Nature. 2012. 483(7391). 603–607.
3) Chazelle B., Lee D.T., On a circle placement problem // Computing. 1986. 36(1-2). 1–16.
4) Ganter B., Kuznetsov S.O., Pattern Structures and Their Projections // preprint MATH-AL-14-2000, Technische Universit at Dresden, Herausgeber, Der Rektor, November 2000.