И. Д. Стаселько, Т. Д. Позняков
(*БГУИР, Минск*)

СРАВНЕНИЕ АГЛОМЕРАЦИИ И РАЗДЕЛИТЕЛЬНОЙ КЛАСТЕРИЗАЦИИ

Одной из широко используемых методик кластеризации является *разделительная кластеризация*, в соответствии с которой для выборки данных, содержащей n записей, задаётся число кластеров k, которое

должно быть сформировано. Затем алгоритм разбивает все объекты выборки на k групп ($k<n$), которые и представляют собой кластеры.

Разделительная кластеризация является более сложной по сравнению с агломерационной кластеризацией, так как в случае разделительной кластеризации нам необходим метод плоской кластеризации в качестве «подпрограммы» для разделения каждого кластера до тех пор, пока у каждого из нас не будет своего собственного одноэлементного кластера. Разделительная кластеризация более эффективна, если мы не создадим полную иерархию вплоть до отдельных листов данных [1]. Временная сложность наивной агломерационной кластеризации составляет $O(n^3)$, поскольку мы тщательно сканируем матрицу $N \times N$ dist_mat на предмет наименьшего расстояния в каждой из N-1 итераций. Используя структуру данных очереди приоритетов, мы можем уменьшить эту сложность до $O(n^2 \log(n))$. Используя еще пару оптимизаций, он может быть уменьшен до $O(n^2)$. Принимая во внимание, что для разделяющей кластеризации при фиксированном количестве верхних уровней, используя эффективный плоский алгоритм, такой как K-Means, делительные алгоритмы линейны по числу шаблонов и кластеров.

По итогу этой работы мы можем увидеть, что в некотором смысле алгоритмы разделения более эффективны. На каждом временном шаге алгоритму нужно только разбить каждый кластер на два таким образом, чтобы он удовлетворял некоторым критериям, например, минимизации ошибки суммы квадратов.

Литература

1 Ломакина, Л.С. Иерархическая кластеризация текстовых документов / Л. С. Ломакина, В. Б. Родионов, А. С. Суркова // Системы управления и информационные технологии. – 2012. – № 2 (48). – С. 39-44.