

С.А. Ермоченко

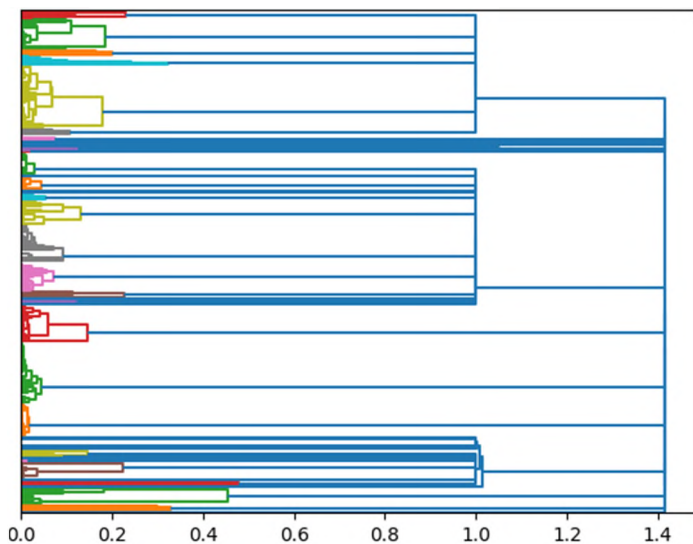
*Витебский государственный университет им. П.М. Машерова,
г. Витебск, Беларусь
SPIN-код: 4105-3911*

РАСПРЕДЕЛЁННЫЕ ВЫЧИСЛЕНИЯ В РЕАЛИЗАЦИИ ИЕРАРХИЧЕСКИХ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Иерархическая кластеризация относится к одному из методов машинного обучения без учителя [1] и может применяться на первом этапе анализа данных некоторой предметной области. При этом данные методы кластеризации позволяют выделять группы схожих объектов с разной степенью близости. В данной работе будет рассмотрен метод агломеративной иерархической кластеризации, основанный на постепенном объединении наиболее близких между собой объектов в группы, которые и называются кластерами. Процесс укрупнения кластеров продолжается до тех пор, пока на последнем этапе все объекты не будут объединены в один общий кластер. Достоинство такой иерархической кластеризации заключается в том, что на начальном этапе анализа, когда исследователю ещё не известна специфика информации об имеющихся объектах, анализ полученной иерархии кластеров позволяет автоматически получать некоторую обобщённую информацию о предметной области. В то время как, например, популярный метод кластеризации К-средних [2] хоть и работает быстрее и использует меньший объём памяти, но требует указания желаемого количества кластеров и распределяет объекты по строго указанному количеству кластеров, что требует от исследователя понимания некоторой специфики предметной области. Также качество кластеризации в алгоритме k-средних зависит от выбора начальных центров кластеров, что тоже требует от исследователя понимания специфики имеющихся данных.

Применение иерархического кластерного анализа является удобным способом начального анализа специфики неизвестной

предметной области, но у этого метода есть ряд своих ограничений. Для анализа данных об n объектах необходимо сформировать матрицу расстояний между всевозможными парами этих объектов, что потребует хранения $n(n - 1) / 2$ чисел в памяти. Таким образом алгоритм имеет квадратичную сложность по памяти и линейную сложность по количеству операций. В связи с чем применение методов для обработки больших объёмов данных (при $n > 10\,000$) может быть затруднительным. Ещё больше процесс анализа может затрудняться способом представления результатов кластеризации. Как правило при иерархической кластеризации строится дендрограмма, показывающая иерархию выявленных кластеров. На рисунке приведён пример дендрограммы, на которой отображено около 150 объектов. Дендрограмма построена с помощью библиотеки `scipy.cluster.hierarchy` языка программирования Python. Как видно, уже на таком количестве объектов информативность диаграммы становится низкой и даже определить количество кластеров верхнего уровня становится сложно.



Дендрограмма, демонстрирующая анализ структуры учебной нагрузки кафедры прикладного и системного программирования ВГУ имени П. М. Машерова

В рамках проведённого исследования для применения кластерного анализа на больших объёмах данных строилась распределённая вычислительная система, позволяющая без хранения матрицы расстояний (вместо этого каждый раз проводится повторный расчёт расстояния между парами объектов) провести процесс кластеризации с сохранением результатов кластеризации в документно-ориентированную базу данных (использовалась СУБД MongoDB), при этом на каждом этапе кластеризации поиск ближайших кластеров с вычислением минимального расстояния между всеми парами уже имеющихся кластеров выполнялся на небольшой ГРИД-системе из 12 персональных компьютеров.

Для представления результатов кластеризации была построена дендрограмма с помощью библиотеки, разработанной на языке программирования JavaScript, что позволяет легко интегрировать её в web-приложения. Данная библиотека позволяет строить интерактивную дендрограмму, которая изначально отображает лишь несколько первых уровней иерархии кластеров и разворачивать по мере необходимости остальные уровни иерархии. При этом библиотека позволяет при разворачивании некоторой ветки дендрограммы динамически подгружать необходимые данные с сервера, на котором проводилась сама кластеризация, что позволяет гибко управлять объёмом используемой на клиентской машине памяти и в целом повысить удобство процесса анализа данных с помощью кластерных методов.

Литература

1. Жамбю М. Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988. 345 с.
2. Arthur D., Vassilvitskii S. How Slow is the k-means Method? // Proceedings of the 2006 Symposium on Computational Geometry (SoCG), June 5–7, 2006. Sedona, Arizona, USA, 2006.