САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
VIII Конгресс молодых учёных

УДК: 004.852

Название тезиса доклада:
«Алгоритм рекурсивной иерархической кластеризации»

Авторы:
Волков Ю.В., Университет ИТМО, Санкт-Петербург

Научный руководитель:
Путин Е.О., Университет ИТМО, Санкт-Петербург

В связи с быстрым развитием области искусственного интеллекта всё больше компаний, исследовательских групп и других организаций уделяют большое внимание сбору данных, с которыми они напрямую или косвенно связаны. Бóльшая часть данных зачастую представляет собой слабоструктурированный набор записей, требующий автоматизированной обработки для дальнейшего принятия решений на его основе.

Изучением алгоритмов и технологий по извлечению полезных знаний из слабоструктурированных данных занимается область машинного обучения, называемая “обучение без учителя”, в частности кластеризация. Однако во многих случаях объёмы данных, подлежащих анализу, оказываются столь большими, что без определённых модификаций алгоритмы кластеризации оказываются абсолютно неприменимы на практике.

Целью проведённой работы является разработка иерархического алгоритма кластеризации, который будет способен кластеризовать большой набор данных, учитывая ограничение по времени и доступной памяти. В рамках данной работы для достижения поставленной цели были сформулированы следующие задачи:

1. Анализ существующих алгоритмов кластеризации
2. Выявление слабых мест существующих алгоритмов
3. Разработка модификации алгоритма иерархической кластеризации с учётом ограничений по времени и памяти.

Основой разработанного алгоритма является разбиение исходного датасета на различные участки, каждый из которых может быть кластеризован отдельно друг от друга. Затем результаты отдельных кластеризаций могут быть обьединены вместе и кластеризованы во второй раз. Было предложено два способа разбиения исходного датасета, а также два способа обьединения отдельно кластеризованных частей. Каждая из четырёх комбинаций имеет свои преимущества и недостаки. Таким образом был разработан не только отдельный алгоритм, но целое семейство алгоритмов, названное “Рекурсивной иерархической кластеризацией” и позволяющее кластеризовать десятки миллионов обьектов быстрее аналогов и эффективнее по памяти, что было подтверждено множественными экспериментами.

Список основных литературных источников:

1. Embrechts, M & Gatti, Christopher & Linton, Jonathan & Roysam, Badrinath. (2013). Hierarchical Clustering for Large Data Sets. 10.1007/978-3-642-28696-4_8.

Автор: Волков Ю.В.
Научный руководитель: Путин Е.О.