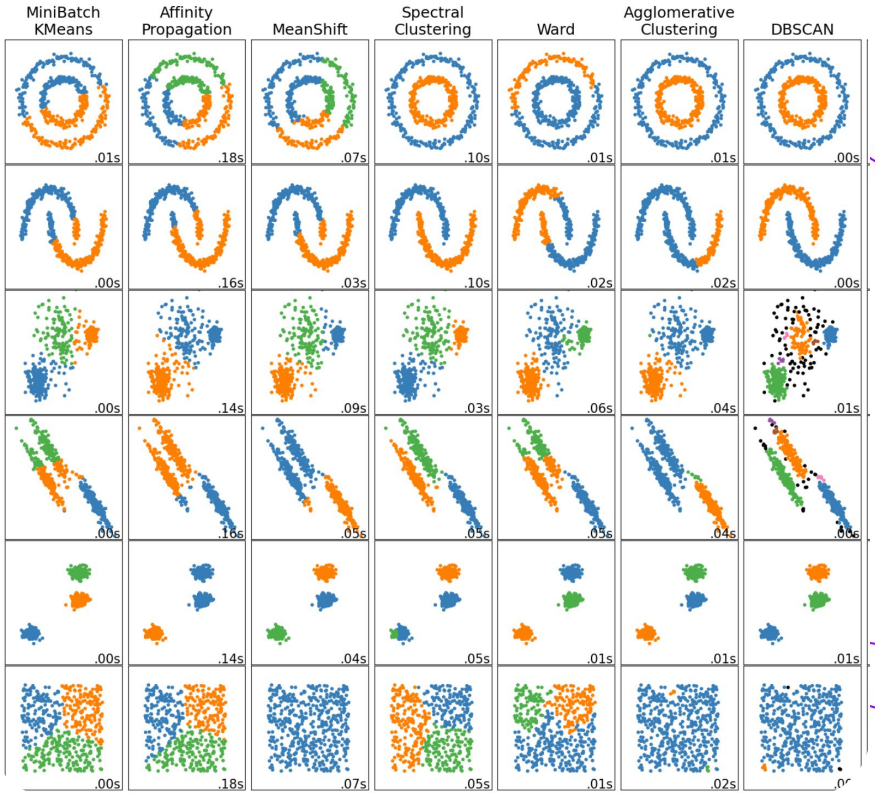
GaMAC: открытая библиотека для автоматического решения задачи кластеризации на GPU

Усов Иван, Кулин Никита, Муратов Симар
parallaxel@yandex.ru, kylin98@list.ru, SYMuratov@itmo.ru

08.10.2024

Задача кластеризации



- Unsupervised разбиение набора данных на непересекающиеся подмножества
- Множество эвристических алгоритмов кластеризации
- Огромное пространство гиперпараметров
- Нет универсальной функции оценки качества разбиений

Существующие реализации

ИТМО

Для задачи кластеризации:

- Scikit-learn (https://scikit-learn.org/stable/modules/clustering.html)
- PyClustering (https://pyclustering.github.io/docs/0.8.2/html/index.html)
- Spark MLLib (https://spark.apache.org/docs/latest/ml-clustering.html)
- CuML (https://docs.rapids.ai/api/cuml/stable/api/#clustering)

Для задач обучения с учителем:

- Auto-Sklearn (https://automl.github.io/auto-sklearn/master/)
- AutoWeka (https://www.automl.org/automl-for-x/tabular-data/autoweka/)
- LightAutoML (https://github.com/sb-ai-lab/LightAutoML)
- FEDOT (https://github.com/aimclub/FEDOT)

Sparkling

Фичи Sparkling:

- Поддержка мультимодальных распределенных данных
- Автоматическая система настройки алгоритмов кластеризации

Ограничения Sparking:

- Страдает производительность на небольших наборах данных
- Сложная процедура развертывания и поддержки окружения
- Небольшой выбор алгоритмов кластеризации и мер качества

GaMAC

Новые фичи в GaMAC:

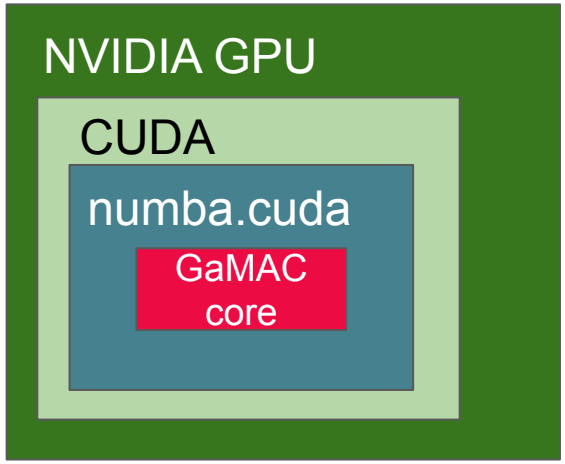
- Буст производительности за счет реализаций на GPU
- Простое использование на обычной рабочей станции
- Разнообразные алгоритмы кластеризации и меры качества
- Пересмотр подхода к обработке мультимодальных данных
- Система мета-обучения для рекомендации мер качества
- Нормальный CI/CD

Аналоги на GPU

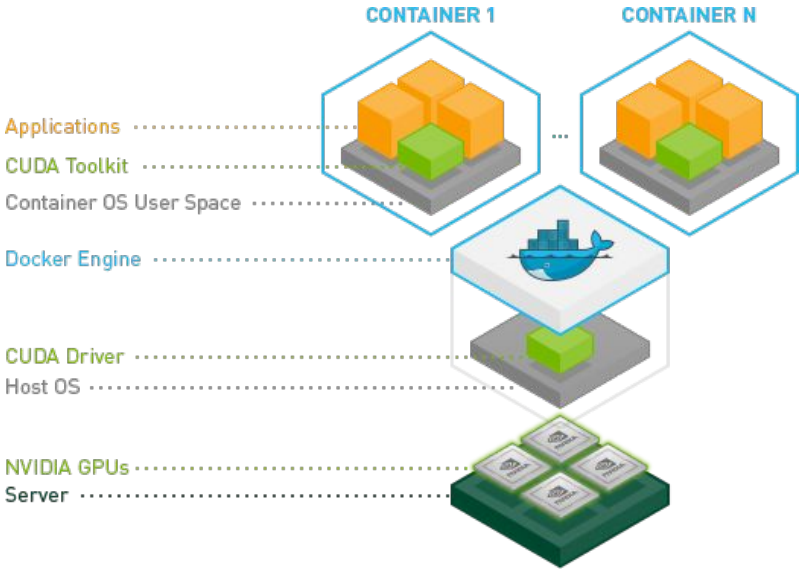
	Кластеризация	Мультимодальность	Мета обучение	AutoML
GaMAC	+	+	+	+
LAMA GPU	-	-	-	+
cuML	+	-	+	-
	Multi-node	SaaS	CI/CD to cloud	MLOps features
GaMAC	+	+	+	+
LAMA GPU	-	-	-	+
cuML	+	+	+	+

Перевод на GPU

ИТМО



GaMAC GPU Dev



GaMAC sandbox

Сведение мультимодальных данных к табличным

Задача: переводим картинки и текст в эмбеддинги для склейки с табличными данными

Гипотезы:

1) Каждую модальность кодировать отдельными моделями
2) Contrastive image-text модели
3) Fusion модели (Image-text to text)

Эксперимент:

1. Получить эмбеддинги картинки и текста
2. Сделать кластеризацию на эмбеддингах
3. Посчитать метрики по кластеризации

Датасеты:

1. MSCOCO
2. EMNIST
3. CIFAR

Кандидаты:

1. E5 + ResNet50
2. E5 + Swin Transformer
3. CLIP-ViT-B
4. CLIP-ViT-L/14
5. Llama-3.2-11B-Vision (In Progress)

Сведение мультимодальных данных к табличным

dataset	method	silhouette	calinski harabasz	davies bouldin	normalized mutual info
EMNIST	clip_b	0.164941	960.231015	2.000735	0.838490
	baseline	0.083806	650.689334	2.633190	0.490402
	swin_e5	0.109977	650.949920	2.339482	0.580193
	clip_l	**0.289205**	**1934.123559**	**1.520208**	**0.983215**
CIFAR	clip_l	**0.317251**	**191.928211**	**1.474333**	**0.993279**
	clip_b	0.255371	163.843154	1.747091	0.984153
	baseline	0.031173	48.950943	3.452677	0.533847
	swin_e5	0.054847	41.406548	3.329345	0.804645
MSCOCO	clip_l	0.050994	106.564796	3.801835	-
	clip_b	0.063904	131.452354	3.271781	-
	baseline	**0.442548**	**431.290599**	3.677789	-
	swin_e5	0.348317	76.934665	**3.258689**	-

CLIP модели лучше работают по сравнению с отдельным кодированием

Рекомендация меры качества

Параметры обучения:

- 1000 двумерных представлений наборов данных
- 200-мерное мета-признаковое описание наборов данных
- ≈ 20 внутренних мер качества
- ≈ 100 асессоров для визуальной оценки разбиений
- Построение полного порядка на 15 разбиениях
- Сравнение ранжирования асессора с ранжированием по мере

ИТМО

GaMAC

https://github.com/parallaxel/GaMAC

Sparkling

https://gitlab.com/rainifmo/sparkling

Спасибо
за внимание!

ITMO more than a UNIVERSITY