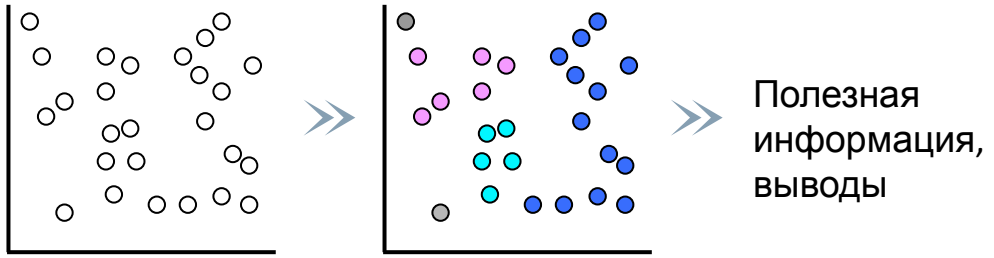
Sparkling: открытая библиотека для автоматического решения задачи кластеризации табличных и мультимодальных данных

Шпинева Полина
polina.shpineva@itmo.ru

06.12.2023

Кластеризация — задача разбиения заданной выборки объектов на непересекающиеся подмножества (кластеры), так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.



Задача *выбора* и *настройки* алгоритма кластеризации на сегодняшний день является **экспертной**

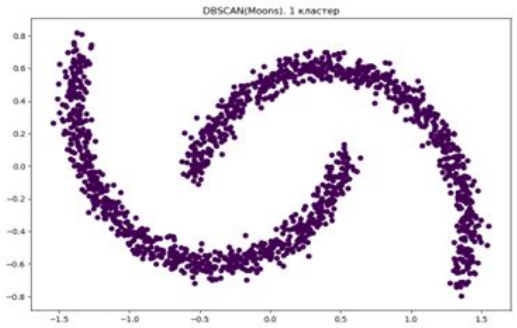
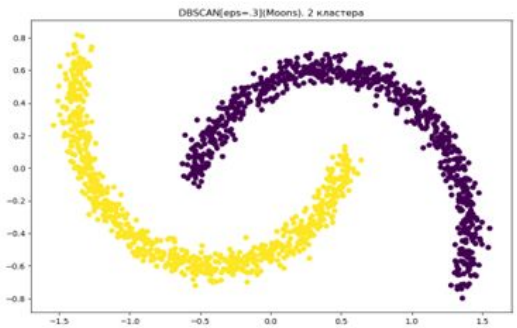
Сферы применения:

- Археология
- Медицина
- Компьютерные науки
- Бизнес-задачи
- Социология
- Лингвистика
- Маркетинг
- Обработка изображений
- Геология
- и тд

Гиперпараметры

Пример: алгоритм DBSCAN и его гиперпараметры:

- радиус шара «плотности», выстраиваемого около каждого элемента;
- минимальное число элементов, входящих в радиусе;

С гиперпараметрами по умолчанию	С настроенными гиперпараметрами
	

Как оценивать разбиения?

Внешние меры

- Функции оценки, которые используют сторонние данные о решаемой задаче, например разметка набора данных на классы.

Внутренние меры (*CVI* — clustering validity index)

- Функции оценки, которые используют данные о структуре самого разбиения.
- Существует более 30 таких мер, среди которых не установлено универсальной.

Существующие реализации

Для задач обучения с учителем:

- Библиотека TPOT
- Библиотека Auto-Sklearn
- Библиотека AutoWeka
- Библиотека HyperOpt

Для задачи кластеризации:

- AutoCluster выбор алгоритма на основе мета-обучения

Проблема:
Для задач кластеризации **не существует** инструмента для автоматической настройки и выбора соответствующего задаче алгоритма, который поддерживал бы мультимодальность и работу с большими наборами данных.

Задачи:

1. Рекомендация меры оценки качества задачи кластеризации
2. Выбор и настройка алгоритма кластеризации

Мета-обучение

ИТМО

Мета-обучение
Перенос знаний о решении одних задач на ускорение поиска решения других задач

Мета-модели
Алгоритмы машинного обучения применяются к **мета-данным** предыдущих экспериментов машинного обучения

Мета-признак описывает свойство задачи.

Примеры:

- Разреженность набора данных
- Число категориальных признаков объектов в наборе данных
- Число возможных меток
- Размер набора данных

Мета-признаковое описание задается вещественными скалярными значениями.

Рекомендация меры оценки качества

Этап обучения:

1. Для каждого набора данных из OpenML:
   a. вычисление 19 мета-признаков
   b. определение лучшей меры качества для каждого
2. Формирование набора данных *DatasetOfDatasets*, где в качестве целевого признака выступают меры качества.
3. Обучение классификатора *CVI_Predictor* на наборе данных из предыдущего пункта

Этап рекомендации для нового набора данных D:

1. Вычисление 19 мета-признаков для набора данных D
2. Формирование вектора, являющегося новой записью для *CVI_Predictor* (по аналогии с наборами из *DatasetOfDatasets*).
3. Предсказание меры качества при помощи обученного мета-классификатора *CVI_Predictor*

Выбранные внутренние меры качества:

- Индекс Калински-Харабаса
- Обобщенный индекс Данна
- Силуэтный индекс
- Мера Дэвиса Болдина

Классификатор	
K Nearest Neighbors	0,58
Random Forest	0,83
Decision Tree	0,79
Naïve Bayesian	0,68
XGBoost Classifier	**0,86**

Выбор и настройка алгоритма кластеризации

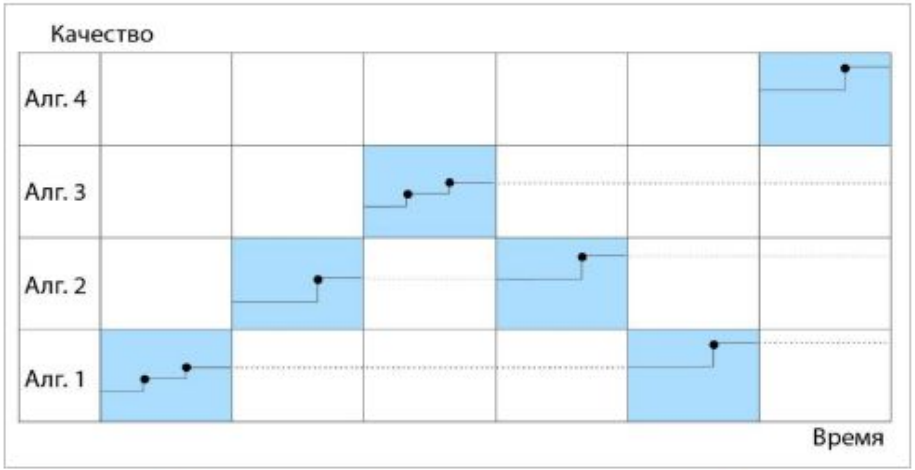
Дано **множество алгоритмов** с соответствующим пространствами гиперпараметров, фиксированное **время** T для поиска оптимального алгоритма.

Необходимо найти компромисс между двумя крайними случаями (*exploration-exploitation tradeoff*):

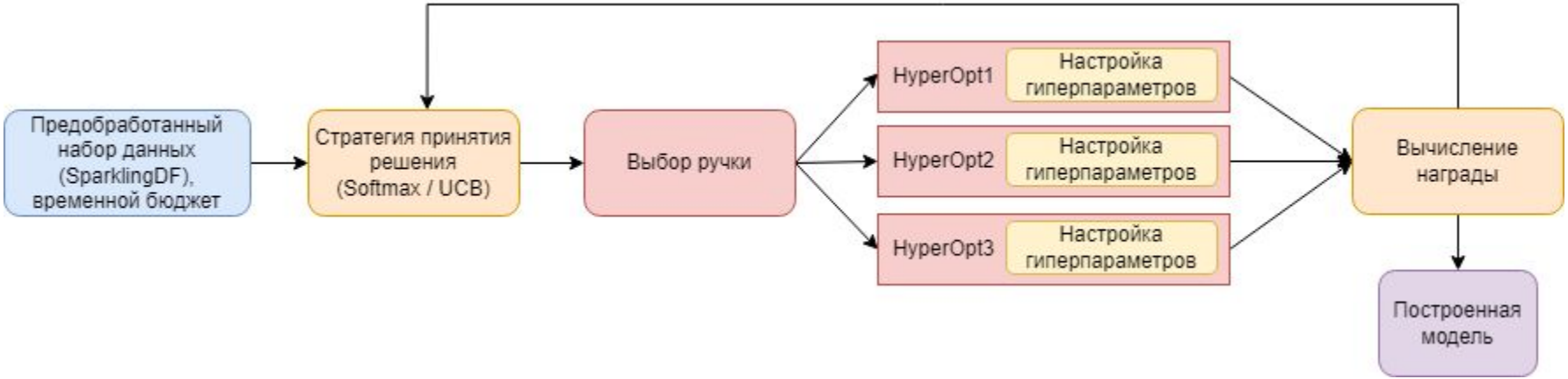
1. **Однородное разделение временного бюджета** между алгоритмами может привести к трате значительной части времени на неэффективные алгоритмы.
2. **Предоставление приоритетов алгоритмам** может привести к потере информации о качестве работы других алгоритмов.

Схема работы алгоритма настройки моделей

Алгоритм на основе обучения с подкреплением, а именно на основе решения *задачи о* ***многоруком бандите.***



Агент итеративно нажимает на различные *ручки*, получает *награду* после каждой итерации. Цель агента – разработать стратегию последовательности активации ручек для максимизации награды, а в общем случае для оптимизации заданной агенту целевой функции.


Предобработанный набор данных (SparklingDF), временной бюджет
Стратегия принятия решения (Softmax / UCB)
Выбор ручки
HyperOpt1
Настройка гиперпараметров
HyperOpt2
Настройка гиперпараметров
HyperOpt3
Настройка гиперпараметров
Вычисление награды
Построенная модель

Модули библиотеки

ИТМО

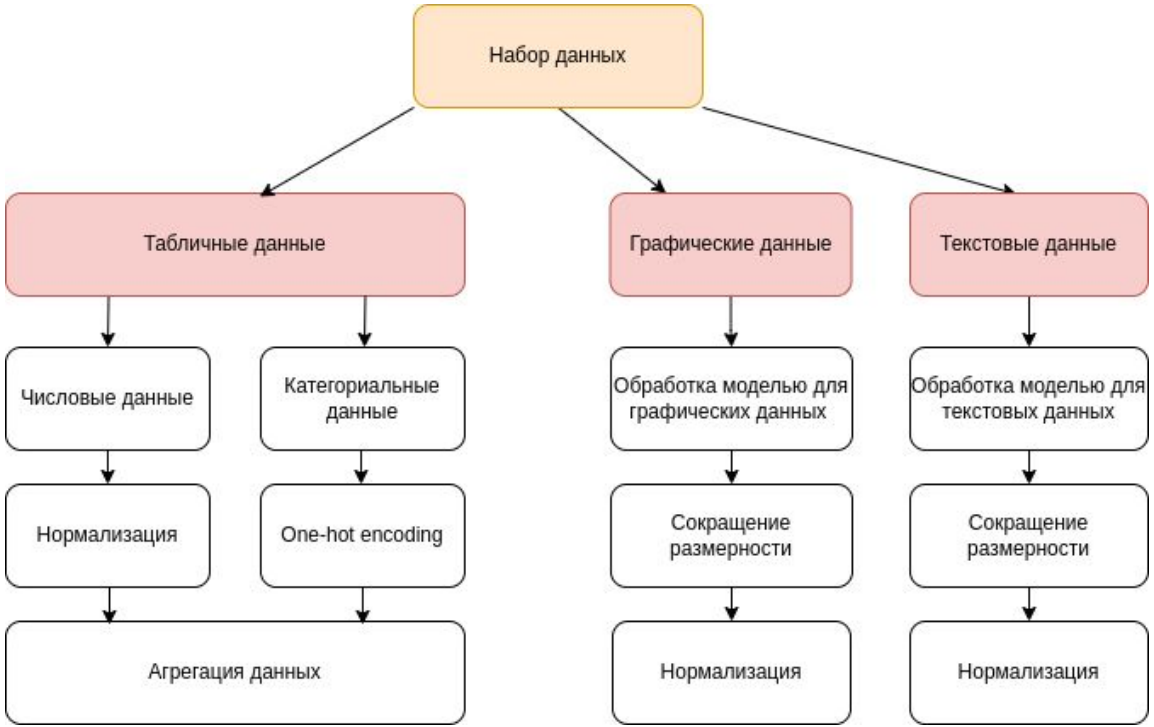
Sparkling

- Препроцессинг, процесс оптимизации и взаимодействие с пользователем.
- Необходимы сторонние модули python в зависимости от конкретной задачи.

Heaven

- Реализации алгоритмов и мер качества.
- Написаны на Scala 2.11.12.
- Не требуют дополнительных зависимостей, помимо Apache Spark.

Мультимодальные данные



- Различные комбинации модальностей:
  - Текст
  - Изображения
  - Табличные данные
- Поддержка двух фреймворков:
  - Pytorch
  - Tensorflow
- Препроцессинг производится при помощи различных моделей HuggingFace.

Подсчет расстояний между модальностями (Heaven)

$$L = \sqrt{\sum_{i=1}^{N} w_i (A_i - B_i)^2}$$

В библиотеке реализованы метрики расстояния:

- Косинусное расстояние
- Канберрское расстояние
- Расстояние Минковского

N – количество модальностей;

w_i – некоторые нормировочные коэффициенты;

$(A_i - B_i)$ – расстояния между модальностями объектов A и B.

Распределенные алгоритмы кластеризации (Heaven)

ИТМО

- K-Means
- MeanShift
- DBSCAN
- Спектральный алгоритм
- BIRCH
- Bisecting K-Means

Spark обрабатывает данные в парадигме MapReduce.

Рекомендация меры качества для реальных наборов данных (Sparkling)

ИТМО

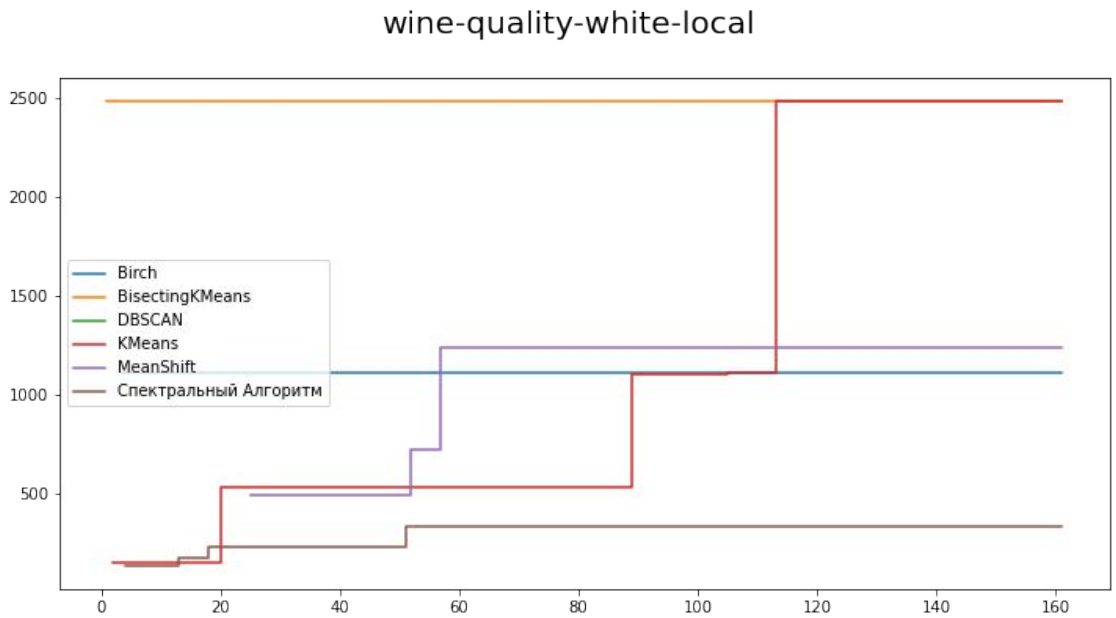
Набор данных	Время, с	Рекомендованная мера
arrhythmia	26	GD41
balance-scale	21	CALINSKI_HARABASZ
cpu	7	GD41
dermatology	17	CALINSKI_HARABASZ
ecoli	10	SCORE Function
german	33	GD41
glass	7	SCORE Function
haberman	8	SCORE Function
heart-statlog	7	CALINSKI_HARABASZ
iono	10	SCORE Function

Результат работы библиотеки на мультимодальных наборах данных. (Sparkling)

Внутренняя мера	Лучший алгоритм по выбранной мере	Время работы, мин
Мера Дэвиса Болдина	Birch	33
Мера Calinski-Harabasz	Birch	35
Мера Данна	MeanShift	66
Силуэтный индекс	Birch	34
Score function	BisectingKMeans	36

Flick dataset – набор данных, содержащий изображения различных ситуаций из жизни людей и по пять предложений-описаний.

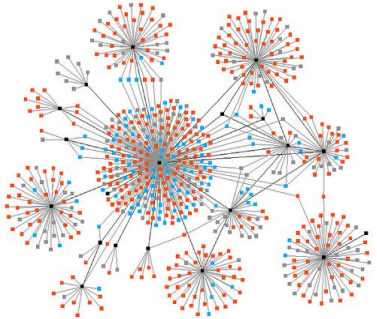
Процесс настройки



Работа алгоритмов на наборе данных, описывающих белое вино; кумулятивный минимум меры Calinski-Harabasz; временной ресурс 20 минут; оптимизатор Optuna; стратегия выбора алгоритма Softmax.

Кейсы реального применения

- Рекомендательная система научных сотрудников для абитуриентов университета ИТМО.



- Построение разбиения археологических артефактов в регионе раскопок *Изюк* (Тюменская область)

Sparkling

https://gitlab.com/rainifmo/sparkling

Спасибо
за внимание!

ITMO more than a UNIVERSITY