ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ ТАБЛИЧНЫХ НАБОРОВ ДАННЫХ КАК ОБЪЕКТОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ

Подготовил: Артюх Севастьян Витальевич, ФИТ-191

Задача кластеризации

В хранилищах наборов данных для кластеризации датасетов чаще используется текстовое описание. Датасеты представляют собой документы, разделенные по темам.

Структура самих данных, характер и распределение их значений в этой задаче не учитывается. При поиске похожих датасетов приходится проводить первичный анализ через графическое представление.

Признаковое пространство

Для вещественных характеристик вычислены: среднее, минимум, максимум, 25, 50 и 75% квантили.

- feature_distance_mean_eucl_N – среднее евклидово расстояние каждой пары столбцов, нормированное по строчке;
- feature_cov – абсолютные значения верхнего треугольника матрицы ковариации без диагонали;
- feature_cov_self – диагональ матрицы ковариации.

Для категориальных характеристик вычислены: минимум и максимум в процентном соотношении от количества записей.

- cat_count – количество категориальных значений каждого столбца;
- cat_unique – количество уникальных категориальных значений каждого столбца;
- cat_freq – частота появления каждого категориального значения в столбце.

Процентное соотношение количества дублирующих, пропущенных строчек к количеству всех строчек никак не повлияли на полученные кластеры.

Проверка гипотезы о тенденции к кластеризации

Пусть X будет набором n точек данных, H – нулевая гипотеза о распределении данных через равные промежутки.

Рассмотрим случайную выборку (без замены) из $m < n$ точек данных с элементами x_i. (Lawson and Jurs (1990)) предлагают выбирать 5% точек данных, чтобы расстояния до ближайших соседей были независимыми и, таким образом, аппроксимировали бета-распределение.

Необходимо:

1) Сгенерировать множество Y из m равномерно распределенных точек;
2) Определить две меры расстояния:
   u_i расстояние $y_i \in Y$ до ближайшего соседа в X и
   w_i расстояние $x_i \in X$ до ближайшего соседа в X.

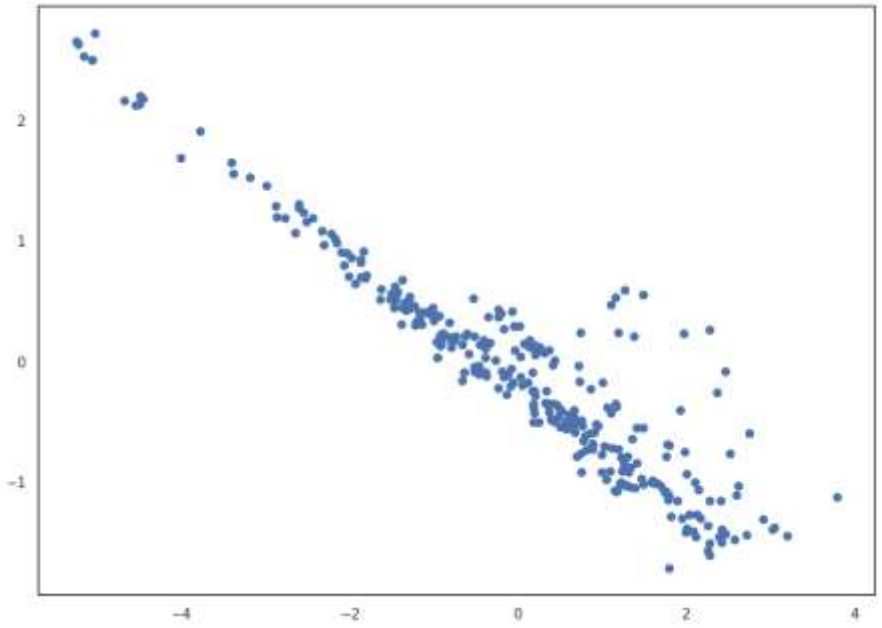
Так как пространство многомерное, то статистика Хопкинса определяется так:

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} \approx 1$$

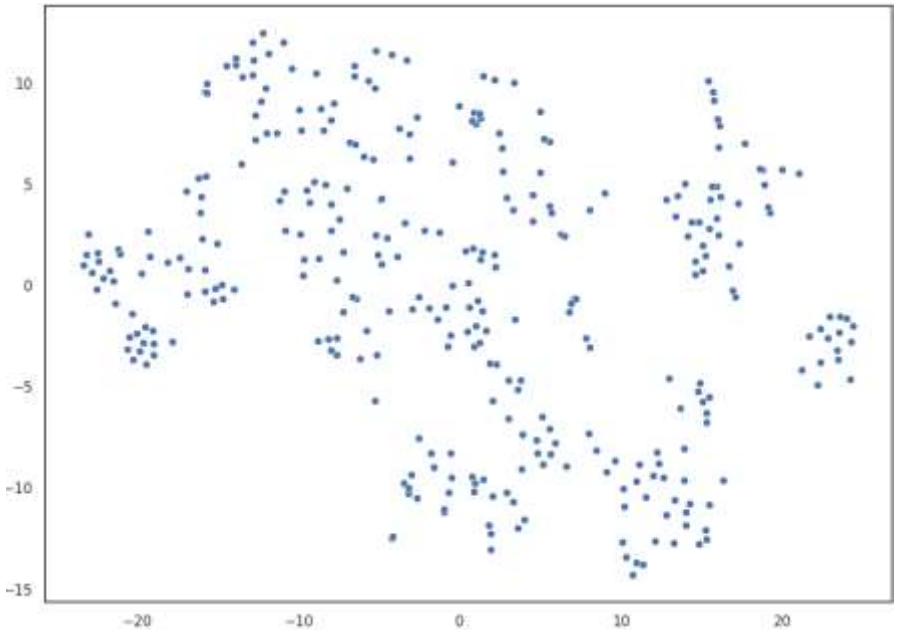
При нормированном X : $H \approx 0.93$

Понижение размерности

PCA



t-SNE



323 датасета

Оценка количества кластеров

Статистика	Лучшее значение	Оптимальное количество кластеров
Коэффициент силуэта	0.26	21
Индекс Calinski–Harabasz	95	14
Индекс Davies–Bouldin	0.8	10

$$-1 \leq Sil(C) \leq 1$$

Или критерий отношения дисперсии

Сравнивает расстояние между кластерами с размером самих кластеров

Оценка количества кластеров Gap-статистикой

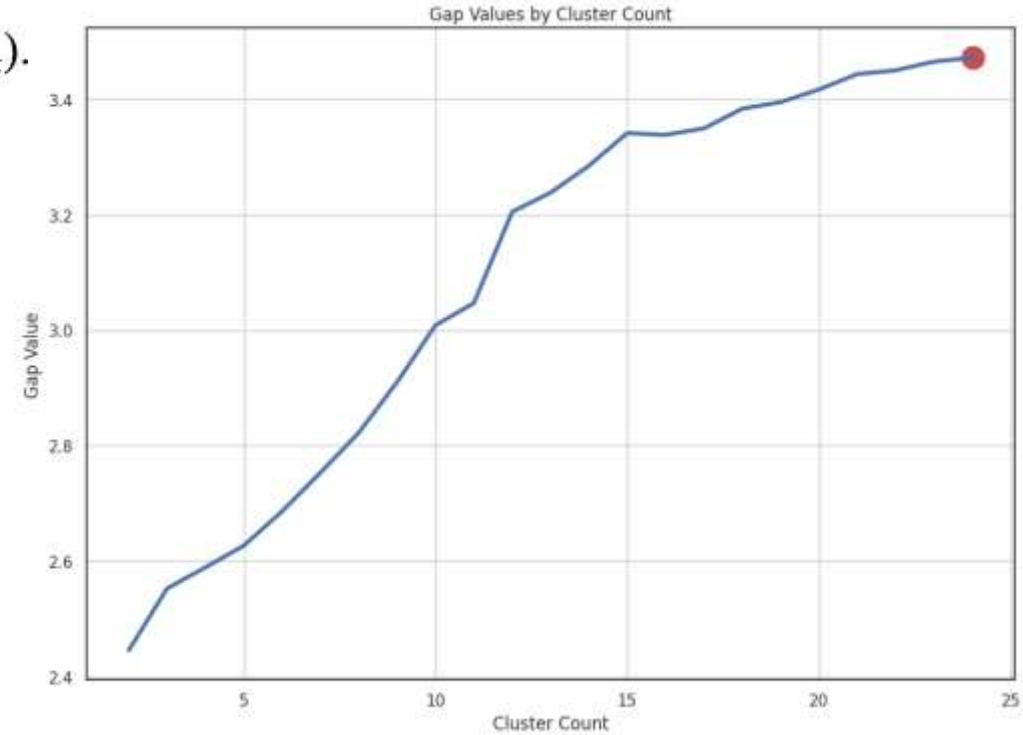
$$\mathrm{Gap}(k) = (1/B)\sum_{b} \log(W^*_{kb}) - \log(W_k).$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \quad D_r = \sum_{i,i' \in C_r} d_{ii'}$$

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

$$s_k = \mathrm{sd}_k \sqrt{(1+1/B)},$$

where $\mathrm{sd}_{k'}$ – denotes the standard deviation of the B Monte-Carlo replicates $\log(W^*_{k'})$

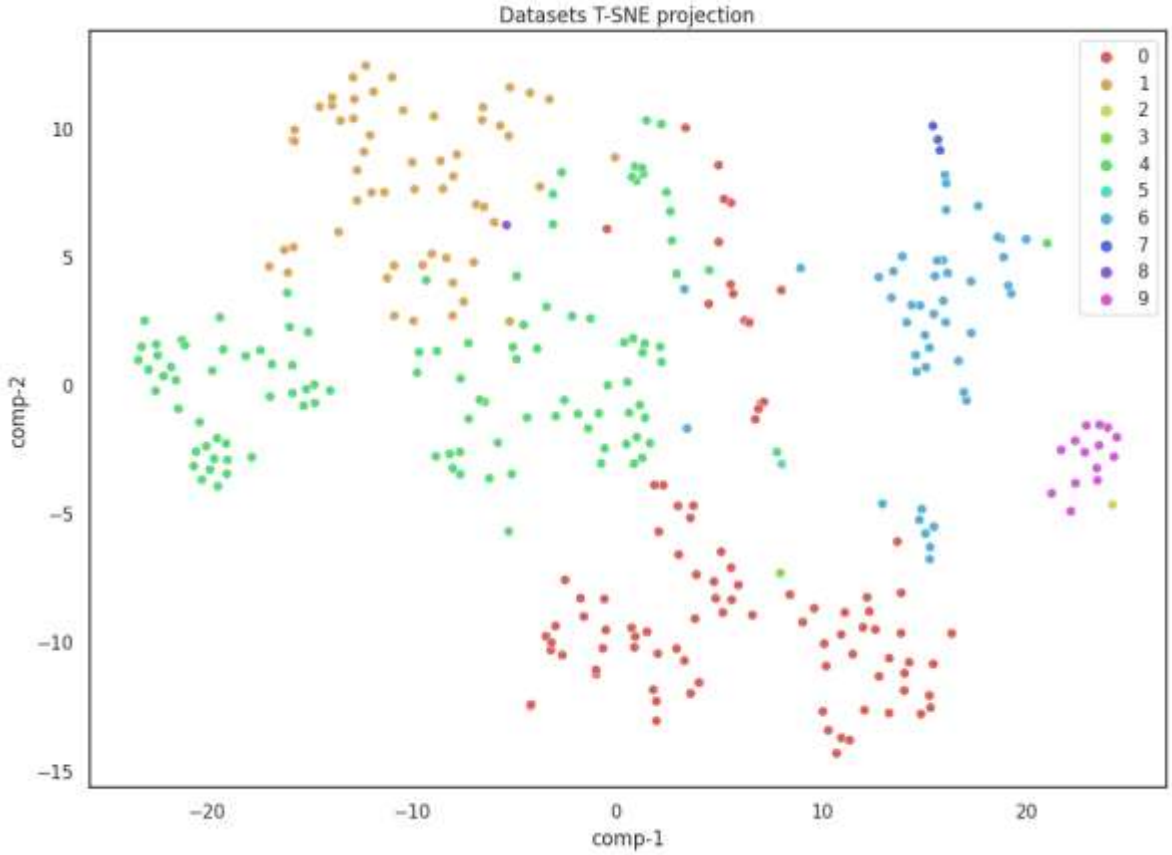


Estimating the number of clusters in a data set via the gap statistic

Оптимальное k = 3, 10, 12, 15, 21

Кластеры на проекции t-SNE

10



Анализ кластеров

Кластер 1

```
rray(['2dplanes', 'BNG(breast-w)', 'Avocado-Prices', 'BNG(stock)',
       'Credit-Card-Dataset-for-Clustering',
       'Default-of-Credit-Card-Clients-Dataset',
       'Eighty-years-of-Canadian-climate-data',
       'Higgs_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'Intersectional-Bias-Assessment', 'Ishwar', 'JapaneseVowels',
       'MAGIC-Gamma-Telescope-Dataset', 'MagicTelescope',
       'MiamiHousing2016', 'NASA_PHM2008', 'NASA_PHM2008_1',
       'ParkinsonSpeechDatasetwithMultipleTypesofSoundRecordings',
       'Premier_League_matches', 'VulNoneVul',
       'WaveformDatabaseGenerator', 'analcatdata_halloffame', 'avila',
       'avocado_sales', 'bfi_dataset', 'blocks', 'cold', 'dataset_sales',
       'default-of-credit-card-clients', 'default_credit_card_p', 'dtt',
       'eeg-eye-state', 'electrical-grid-stability', 'fried',
       'grid_stability', 'helena',
       'helena_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'higgs', 'kdd_ipums_la_97-small', 'kin8nm', 'microaggregation2',
       'numerai28.6',
       'numerai28.6_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True'
       'pendigits', 'premier_league_with_tda', 'ringnorm', 'sarcos',
       'segment',
       'segment_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'shill-bidding', 'svmguide3', 'sylvine',
       'sylvine_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'turing_binary', 'twonorm', 'ulaanbaatar-weather-2015-2020',
       'wind'], dtype=object)
```

Кластер 0

```
'bankmarketing', 'beijing-pm2.5', 'bias-correction',
'burst-header-packet', 'california', 'california_housing',
'california_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
'churn',
'churn_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
'colleges_aaup', 'combined-wine-data', 'cpu_act', 'cpu_activity',
'cpu_small', 'datapm2.5', 'delta_elevators', 'diamonds',
'elevators', 'eye_movements',
'eye_movements_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
'garments-worker-productivity', 'gas-turbine-2011', 'hmeq_p',
'house_16H', 'house_8L', 'house_sales', 'house_sales_reduced',
'houses', 'jm1', 'kc1', 'kings_county', 'miami_housing',
'mlr_rpart_rng', 'mv', 'naval_propulsion_plant',
'obesity-level-indicators', 'online-shoppers-intention',
'online_shoppers', 'page-blocks', 'parkinsons-telemonitoring',
'pbcseq', 'pc1', 'physicochemical-protein',
'physiochemical_protein', 'pm25dataset', 'scpf', 'skillcraft1',
'space_ga', 'steel-plates-fault', 'stock_fardamento02', 'test_dsn',
'treasury', 'video_transcoding', 'wine', 'wine-quality-white',
'wine_quality', 'wingstop_stock_prices'], dtype=object)
```

Кластер 4

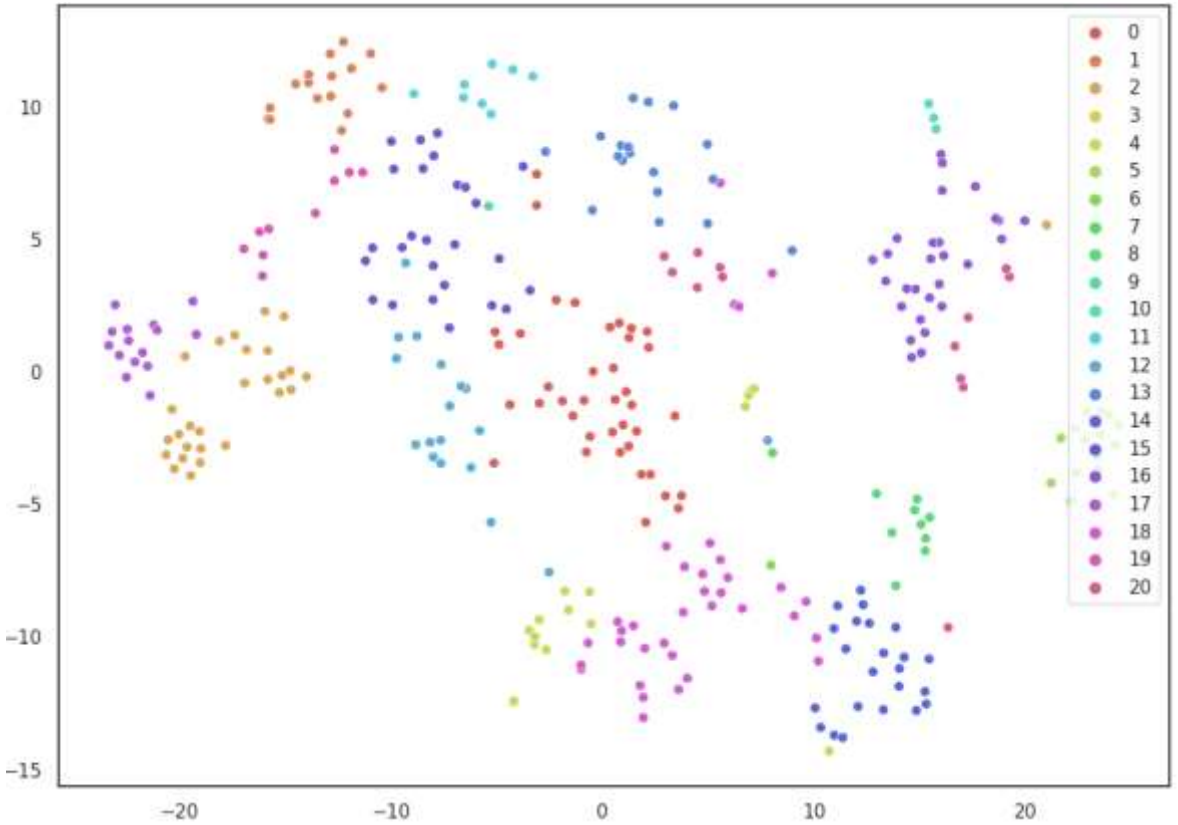
```
'sick', 'sulfur', 'svmguide1', 'thyroid-ann', 'thyroid-dis',
'visualizing_soil', 'volcanoes-a1', 'volcanoes-a2', 'volcanoes-a3',
'volcanoes-a4', 'volcanoes-b1', 'volcanoes-b2', 'volcanoes-b3',
'volcanoes-b4', 'volcanoes-b5', 'volcanoes-b6', 'volcanoes-c1',
'volcanoes-d1', 'volcanoes-d2', 'volcanoes-d3', 'volcanoes-d4',
'volcanoes-e1', 'volcanoes-e2', 'volcanoes-e3', 'volcanoes-e4',
'volcanoes-e5', 'wall-robot-navigation', 'weather_izmir', 'wilt',
'wilt_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
'wine-quality-red', 'yeast'], dtype=object)
```

Анализ кластеров

```
 Кластер 2
array(['Human-Memory-and-Cognition'], dtype=object)
 Кластер 3
array(['Global-Cause-of-the-Deaths-other-than-diseases'], dtype=object)
 Кластер 5
array(['Bitcoin-Stock-Data'], dtype=object)
 Кластер 7
rray(['Apple-(AAPL)-Historical-Stock-Data',
       'Google-Stock-10Year-data2004-2020',
       'Top-100-2020-Cryptocurrency-Daily-Market-Price'], dtype=object)
 Кластер 8
rray(['The-Price-and-Sales-of-Avocado'], dtype=object)
```

Кластеры на проекции t-SNE

21



Анализ кластеров

Кластер 0

```
array(['AI4I2020', 'BNG(lowbwt)', 'Brazilian_houses',
       'Boston-Weather-Data-Jan-2013---Apr-2018',
       'Brazilian_houses_reproduced', 'COVID19-cases-by-country',
       'Case-Study-Applicants-for-a-Gold-Digger-position',
       'Consumer-Price-Index-in-Denver-CO', 'Employee-Turnover-at-TECHCO
       'Football---Expected-Goals-Match-Statistics',
       'Hazardous-Driving-Spots-Around-the-World', 'Medical-Appointment'
       'MembershipWoes', 'Metro-Manila-Flood-Landscape-Data',
       'Myanmar-Air-Quality(2019-to-2020-Oct)',
       'New-Delhi-Rental-Listings', 'Predicting-Critical-Heat-Flux',
       'Solar-Radiation-Prediction', 'UCI_churn', 'ada_prior', 'adult',
       'adult-census',
       'adult_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'airlines_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'bank-marketing',
       'bank-marketing_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_
       'bank8FM', 'bankmarketing', 'beijing-pm2.5', 'credit',
       'credit_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'dataset_time_9', 'eucalyptus', 'mfeat-morphological',
       'national-longitudinal-survey-binary', 'nfl_games', 'shuttle',
       'shuttle_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'stock_fardamento02', 'visualizing_soil'], dtype=object)
```

Кластер 17

```
array(['Bank-Note-Authentication-UCI', 'Daily-Wheat-Price',
       'Dogecoin-Historical-Data', 'Emotions--Sensor-Data-Set',
       'banknote-authentication',
       'phoneme_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'volcanoes-a2', 'volcanoes-a3', 'volcanoes-a4', 'volcanoes-e1',
       'volcanoes-e2', 'volcanoes-e3', 'volcanoes-e4', 'volcanoes-e5',
       'yeast'], dtype=object)
```

Кластер 2

```
rray(['Chernobyl-Air-Concentration', 'Credit-Risk-Dataset',
       'NYC-Hourly-Temperature', 'Run_or_walk_information',
       'artificial-characters', 'brazilian_houses', 'calhousing', 'cpu',
       'mammography', 'phoneme', 'pollen', 'quake', 'rl',
       'rl_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'shapes', 'volcanoes-a1', 'volcanoes-b1', 'volcanoes-b2',
       'volcanoes-b3', 'volcanoes-b4', 'volcanoes-b5', 'volcanoes-b6',
       'volcanoes-c1', 'volcanoes-d1', 'volcanoes-d2', 'volcanoes-d3',
       'volcanoes-d4', 'wall-robot-navigation'], dtype=object)
```

Кластер 15

```
rray(['Avocado-Prices', 'BNG(stock)',
       'Credit-Card-Dataset-for-Clustering',
       'Default-of-Credit-Card-Clients-Dataset',
       'MAGIC-Gamma-Telescope-Dataset', 'MagicTelescope',
       'MagicTelescope_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True'
       'MagicTelescope_seed_1_nrows_2000_nclasses_10_ncols_100_stratify_True'
       'MiamiHousing2016', 'NASA_PHM2008', 'Premier_League_matches',
       'VulNoneVul', 'abalone', 'avocado_sales', 'blocks',
       'eeg-eye-state', 'electrical-grid-stability', 'electricity',
       'electricity_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'grid_stability', 'kdd_ipums_la_97-small', 'microaggregation2',
       'sarcos', 'shill-bidding', 'sylvine',
       'sylvine_seed_0_nrows_2000_nclasses_10_ncols_100_stratify_True',
       'ulaanbaatar-weather-2015-2020'], dtype=object)
```

Кластер 9

```
array(['The-Price-and-Sales-of-Avocado'], dtype=object)
```

Заключение

Выявлена тенденция к кластеризации набора данных. Кластеризовано 323 табличных наборов данных с OpenML.

Есть ряд проблем с содержательной интерпретацией кластеров.

Решение этой задачи вместе с задачей определения тематики позволило бы создавать сводки похожих датасетов в разных темах или одинаковых темах, к тому же по произвольному набору данных всегда получалось бы находить похожий датасет.

Диапазон записей в датасетах от 1 до 250 тысяч, количество столбцов до 30. Также в них присутствовало минимум 1 категориальный и 2 вещественных столбца.

Библиографический список

1. Загоруйко Н. Г. Гипотезы компактности и λ-компактности в методах анализа данных, Сиб. журн. индустр. матем., 1998, том 1, номер 1, С. 114–126.

2. Richard G. Lawson, Peter C. Jurs. New index for clustering tendency and its application to chemical problems // The Journal for Chemical Information and Computer sciences, 1990, 30, 1, 36–41

3. Мюллер Адреас П. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. / П. Мюллер Адреас, Гвидо Сара. – Москва : Вильямс, 2017. – 480 с.

4. Robert Tibshirani, Guenther Walther, Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic // Journal of the Royal Statistical Society. Series B. 2001. Part 2, P. 411-423

Спасибо за внимание!