ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ  
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра АСУ

Отчет

о лабораторной работе №2

по дисциплине «Интеллектуальный анализ данных»

на тему: «Применение библиотеки Scikit-Learn для решения задач кластеризации»

Выполнил:

студент группы ИСТ-19б

Карпович В. Д.

Проверили:

Васяева Т. А.

Шуватова Е. А.

Донецк – 2022

**Цель работы:** изучение алгоритмов кластеризации, изучение возможностей библиотеки Scikit-Learn для решения задач кластеризации.

Порядок выполнения работы

1. Выбрать одну из баз данных (согласовать с преподавателем), список которых представлен на сайте http http://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table. Скачать данные с сайта или можно использовать данные, полученные на практике, результаты НИРС или данные для бакалаврской работы.
2. Исходные данные должны храниться в файле. Программным способом считать эти данные и сформировать обучающую и тестовую выборки.
3. Разработать программу на любом языке программирования с использованием любых существующих библиотек. Рекомендован язык Python с использованием библиотек NumPy, Pandas, Scikit-Learn. Программа должна обеспечивать: реализацию метода кластеризации, который указан в задании; визуализацию результатов; тестирование модели; оценка точности модели; сохранение / считывание модели; вывод модели.
4. Согласно варианту заданий (таблица 2.8) построить несколько моделей. В процессе построения моделей, анализируйте ее качество и записывайте результаты в таблицу, указывая значения гиперпараметров.

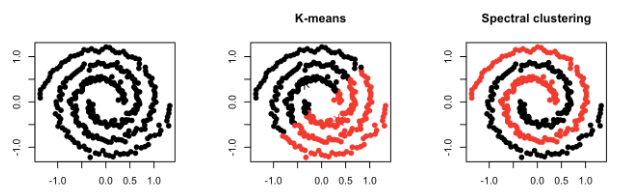
**Индивидуальное задание:** Вариант 4, Spectral clustering

Спектральная кластеризация

Спектральная кластеризация

При спектральной кластеризации точки данных рассматриваются как узлы графа. Затем узлы сопоставляются с низкоразмерным пространством, которое можно легко разделить для формирования кластеров. Спектральная кластеризация использует информацию из собственных значений (спектра) специальных матриц (т.е. матрицы сродства, матрицы степени и матрицы Лапласа), полученных из графика или набора данных.

В таком методе точки, которые соединены или находятся непосредственно рядом друг с другом, помещаются в один и тот же кластер.



Спектральная кластеризация не зависит от количества параметров объектов. Но проблемы могут возникнуть при применении к очень большим наборам данных.

**Описание набора данных**

Данные были извлечены из изображений, снятых с подлинных и поддельных образцов, похожих на банкноты. Для оцифровки использовалась промышленная камера, обычно используемая для проверки печати. Окончательные изображения имеют размер 400x400 пикселей. За счет объектива и расстояния до исследуемого объекта были получены полутоновые изображения с разрешением около 660 dpi. Инструмент Wavelet Transform использовался для извлечения признаков из изображений.

Атрибуты:

1. дисперсия вейвлет-преобразованного изображения (непрерывное)
2. асимметрия вейвлет-преобразованного изображения (непрерывное)
3. эксцесс вейвлет-преобразованного изображения (непрерывный)
4. энтропия изображения (непрерывная)
5. класс (целое число)

Источник:

<https://archive.ics.uci.edu/ml/datasets/banknote+authentication>

**Листинг программы**

import pandas

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

from sklearn.cluster import SpectralClustering

db = pandas.read\_csv('data\_banknote\_authentication.txt')

print(db)

X = db.loc[:, 'variance':'entropy']

Y = db['class']

model = SpectralClustering(n\_clusters = 2, assign\_labels = 'discretize', affinity = 'nearest\_neighbors', n\_neighbors = 10)

model.fit(X, Y)

predict = model.fit\_predict(X)

counter = 0

for i in range(len(predict)):

  if (predict[i] == Y[i]):

    counter += 1

print()

print("Точность предсказания: {:d} \* 100 / {:d} = {:.2f}%".format(counter, len(predict), counter \* 100 / len(predict)))

plt.scatter(db['variance'], db['kurtosis'], c = predict)

plt.show()

**Результат:**

