Головня С.А., магістрант, гр. ПІ-49м,
Марчук Г.В. старш. викладач
Державний університет «Житомирська політехніка»

КЛАСТЕРНИЙ АНАЛІЗ: ІЄРАРХІЧНА КЛАСТЕРИЗАЦІЯ

Кластерний аналіз з'явився у 1939 р. Його запропонував вчений К. Тріон. Дослівно термін "кластер" в перекладі з англійської "cluster" означає гроно, згусток, пучок, група. Кластеризація – це, як правило, метод групування подібних точок даних таким чином, що точки в одній групі більше схожі між собою, ніж точки в інших групах. Група подібних точок даних називається кластером. Сам кластерний аналіз - це не один конкретний алгоритм, а загальна задача, яку потрібно вирішити. Це може бути досягнуто за допомогою різних алгоритмів, які суттєво відрізняються своїм розумінням того, що являє собою кластер та як їх ефективно знайти.

Методи кластерного аналізу можуть бути ієрархічними і неієрархічними. Ієрархічна кластеризація починається з трактування кожного спостереження як окремого кластеру. Потім він неодноразово виконує наступні два кроки: 1- ідентифікує два кластери, які є найближчими один до одного, і 2 - об'єднає два найбільш схожі кластери. Це триває, поки всі кластери не об'єднаються.

Етапи проведення кластерного аналізу:

– Підготовка даних до кластеризації. Виділення даних для аналізу.

```
data_for_clust=data.drop(data.columns[0], axis=1).values
data_for_clust[0]
```

– Виконуємо препроцесинг за допомогою бібліотеки sklearn:

```
from sklearn import preprocessing
dataNorm = preprocessing.scale(data_for_clust)
```

– Вираховуємо відстань між кожним набором даних, вираховуємо евклідову відстань (за замовчуванням).

```
data_dist = pdist(dataNorm, 'euclidean')
```

– Головна функція ієрархічної кластеризації проводить об’єднання елементів в кластери і збереження в спеціальній змінній (використовується далі для візуалізації і виділення кількості кластерів):

```
data_linkage = linkage(data_dist, method='average')
```

– Використаємо метод Ліктя, він дозволяє оцінити оптимальну кількість сегментів. Метод Ліктя - це графічне відображення, рекомендація по кількості кластерів. Коли верхні і нижні лінії сходяться максимально один до одного - саме таку кількість кластерів і рекомендується

розглядати. Як можна побачити по результатах виконання дій - система рекомендує 3 кластери (рис. 1).

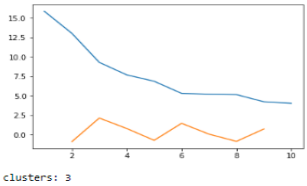

Рис.1. Рекомендована кількість кластерів для розгляду

– Виконуємо побудову дендрограми для наочності результатів кластеризації.

```
nCluster=6
fancy_dendrogram(data_linkage,truncate_mode='lastp',
      p=nCluster,leaf_rotation=90.,leaf_font_size=12.,
show_contracted=True, annotate_above=10, )
plt.show()
```

Основним результатом ієрархічної кластеризації є дендрограма (рис.2), яка показує ієрархічну залежність між кластерами. Виконавши побудову можна зробити висновок про відстань між кластерами, чим вище стовпець тим більше відстань між кластерами, кластеризація виконана у вигляді вкладених груп. В результаті кластеризації в кожний кластер визначена певна група даних, наприклад група з 699 клієнтів, яка є не вкладеною і відрізняється від інших (рис. 2).

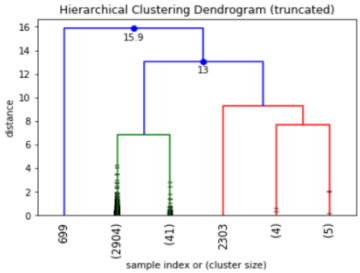

Рис.2. Ієрархічна кластеризація

Ієрархічна кластеризація зазвичай працює шляхом послідовного об'єднання подібних кластерів, як показано вище. Це відоме як агломераційна ієрархічна кластеризація. Теоретично це також можна зробити, спочатку групуючи всі спостереження в один кластер, а потім послідовно розбиваючи ці кластери. Це відоме як розділення ієрархічної кластеризації. Особливість ієрархічної кластеризації, що всі кластери вкладені, тобто слідують один за одним.

УДК 004