# ЛАБОРАТОРНА РОБОТА № 7
# ЗАДАЧА КЛАСТЕРИЗАЦІЇ

## Мета роботи

На практиці вивчити роботу алгоритмів кластеризації, навчитися інтерпретувати результати їх роботи і вибирати найкращий метод для розв'язуваної прикладної задачі.

## Основні теоретичні відомості

У лабораторній роботі розглядаються наступні методи кластеризації (у дужках наведено назву на WEKA):

– поділяючий метод кластеризації K-середніх (*SimpleKMeans*);
– ієрархічний метод кластеризації (*HierarchicalClusterer*)
– імовірнісний метод кластеризації EM (*EM*)
– ієрархічний метод кластеризації COBWEB (*COBWEB*)
– метод заснований на щільності розташування об'єктів DBSCAN (*DBScan*).

Для додання відстутніх методів в WEKA необхідно встановити пакет optics_dbScan.

### *Параметри налаштування алгоритмів*

Розглянемо параметри налаштування використовуваних алгоритмів кластеризації в WEKA (табл. 7.1).

Таблиця 7.1 – Параметри налаштування кластеризаторів

| Метод | Параметры |
|---|---|
| *SimpleKMeans* | displayStdDevs – відображати значення стандартного відхилення для числових атрибутів і підрахунки для номінальних атрибутів.<br>distanceFunction – функція відстані.<br>dontReplaceMissingValues – не замінювати пропущені значення середнім значенням або модою.<br>maxIterations – максимальна кількість ітерацій алгоритму.<br>numClusters – кількість кластерів.<br>preserveInstancesOrder – зберігати порядок примірників у вибірці.<br>seed – випадковий сид для рандомізації вибірки. |

Продовження табл. 7.1.

| Метод | Параметры |
|---|---|
| *Hierarchical Clusterer* | *distanceFunction* – функція відстані.<br>*distanceIsBranchLength* – у дендрограмі висота лінії, що зв'язує кластери, буде показувати відстань між ними.<br>*linkType* – тип зв'язку для розрахунку відстані між двома кластерами.<br>*numClusters* – кількість кластерів.<br>*printNewick* – виводити кластери в форматі Newick. |
| *EM* | *displayModelInOldFormat* – використовувати старий формат представлення моделі (у випадках великої кількості кластерів).<br>*maxIterations* – максимальна кількість ітерацій алгоритма.<br>*minStdDev* – мінімальне значення стандартного відхилення.<br>*numClusters* – кількість кластерів (встановити значень -1 для автоматичного вибору кількості кластерів).<br>seed – випадковий сід для рандомізації вибірки. |
| *DBSCAN* | *database*_Type – використовувана база даних.<br>*database*_distanceType – функція відстані.<br>*epsilon* – радіус пошуку.<br>*minPoints* – мінімальна кількість об'єктів усередині радіуса. |
| *COBWEB* | *acuity* – мінімальне значення стандартного відхилення для числових атрибутів.<br>*cutoff* - встановити поріг до якого відсікати вузли дерева.<br>*saveInstanceData* – зберегти інформацію про примірники для візуалізації.<br>*seed* – випадковий сід для рандомізації вибірки. |

***Інтерпретація результатів кластеризації в WEKA***

Розглянемо результати роботи кластеризаторів в WEKA (*Clusterer output*).

Секція «*Clustering model*» відображає побудовану модель.

Для алгоритму SimpleKMeans ця секція буде містити кількість ітерацій алгоритму, загальну квадратичну помилку для всіх кластерів

та центроїди побудованих кластерів. В ній також буде вказано застосований вид обробки порожніх значень атрибутів у об'єктів.

Для алгоритму EM буде вказана кількість кластерів, на які було розбито дані, кількість ітерацій алгоритму та центроїди побудованих кластерів.

Для алгоритму COBWEB буде вказана кількість об'єднань та розділів даних, кількість побудованих кластерів та текстове представлення побудованої ієрархії.

Для алгоритму DBSCAN ця секція буде містити налаштування алгоритму, кількість побудованих кластерів, віднесення кожного з об'єктів вибірки до конкретного кластеру чи до викидів.

Секція «*Model and evaluation*» містить інформацію про кількісний розподіл екземплярів по кластерах. При цьому буде вказано, скільки об'єктів було кластеризовано (Clustered Instances), а скільки не увійшли в жоден з кластерів (Unclustered instances).

Якщо було обрано опцію «*Classes to clusters evaluation*» (порівняння попередньої заданих класів з кластерами), то ця секція також буде містити результати оцінки якості кластеризації. Буде вказано, який з побудованих кластерів відповідає якому класу, буде побудовано матрицю помилок та вказана кількість невірно кластеризованих екземплярів.

## Завдання на лабораторну роботу

*Частина А*

1. Виконайте наступні завдання для набору даних 'bank.arff'.
2. Запустіть алгоритм кластеризації SimpleKMeans, задаючи значення параметра K (кількість кластерів) від 1 до 12 .
3. Запишіть в таблицю значення сум квадратичних помилок, одержуваних при різних значеннях K. Що означає цей параметр і як змінюються його значення?
4. Для значення K=5 укажіть:
   - скільки кластерів було створено;
   - скільки примірників потрапило в кожен з кластерів (вказати кількість і відсоток);
   - скільки ітерацій знадобилося для кластеризації даних;
   - складіть таблицю з характеристиками центроїдів.
5. Для значення K=5 візуалізуйте результати кластеризації (по осі абсцис відкласти назву (номер) кластера, по осі ординат - номер

примірника в кластері) та дайте оцінку отриманим результатам:

- чи є значна відмінність у значеннях атрибуту «вік» (age) між кластерами?
- у яких кластерах домінують жінки (female), а в яких чоловіки (male)?
- що можна сказати про значення атрибута «регіон» (region) у кожному кластері?
- що можна сказати про розкид значень атрибуту «дохід» (income) між кластерами?
- у яких кластерах домінують сімейні люди (married), а в яких холості (unmarried)?
- у якій кластер потрапило найбільше людей з машинами?
- у яких кластерах переважають люди з ощадними рахунками (savings accounts)?
- що можна сказати про розкид значень атрибуту «поточний банківський рахунок» (current account) між кластерами?
- що можна сказати про розкид значень атрибуту «іпотека» (mortgage holdings) між кластерами?
- які кластери в основному складаються з людей, які придбали PEP (особистий план купівлі акцій), і які з людей, які не придбали його?

6. Запустіть алгоритм кластеризації EM та оцініть результати.

*Частина Б*

7. Виконайте наступні завдання для набору даних ‘iris.arff’

8. Запустіть алгоритм кластеризації SimpleKMeans з K=3 та оцініть якість кластеризації, порівнюючи кластери з попередньо заданими класами:

- запишіть значення суми квадратичних помилок, кількість об'єктів в кластерах та характеристики кожного центроїду;
- проаналізуйте як співвідносяться кластери та значення цільового атрибуту, скільки екземплярів було віднесено до «невірних» кластерів, який клас виявився «складним» для виділення;
- візуалізуйте результати, використовуючи різні атрибути для осі ординат (при візуалізації екземпляри, позначені квадратами були віднесені до «невірного» кластеру);
- визначте, на що впливає параметр «seed» і чому він є важливим при кластеризації методом k-середніх; для

цього проведіть експерименти з різними значеннями параметру і порівняйте отримані результати.

*Частина В. Ієрархічна кластеризація*

9. Завантажте набір даних ‘flagdata.arff. Цей файл представляє атрибути прапорів деяких європейських країн. Виконайте наступні завданні:

- Запустіть алгоритм COBWEB з параметрами C=0,4 (0,35), saveInstanceData = True, cluster mode = Use training set;
- візуалізуйте отриману дендрограму та запишіть її, вкажіть, які країни потрапили в який кластер;
- укажіть, що спільного у прапорів, що опинилися в одному кластері.

10. Завантажте набір даних ‘zoo.arff’ і виконайте наступні завдання:

- оберіть з вибірки частину тварин на власний розсуд (наприклад, ссавців);
- запустіть алгоритм Hierarchical Clusterer (тип тварини не використовувати в кластеризації, а назву за допомогою фільтру перетворити на рядковий тип – NominalToString);
- проекспериментуйте з налаштуванням алгоритму та візуалізуйте результати його роботи;
- оцініть, чи є логічний сенс в створюваних кластерах.

*Частина Г. Алгоритм DBScan*

11. Для використання алгоритму щільністної кластеризації згенеруйте набор даних за допомогою алгоритму BIRCHCluster. В наборі згенеруйте також флаг класу.

12. За допомогою налаштувань методу DBScan досягніть найкращої кластеризації даних.

13. Кластеризуйте набір даних за допомогою інших алгоритмів. Який з алгоритмів виявився найбільш ефективним?

**Контрольні питання**

1. У чому полягає задача кластеризації? Наведіть практичний приклад?

2. Що таке навчання з учителем і без учителя? До якого типу належить задача кластеризації?

3. Задача кластеризації є описовою або прогнозуючою і чому?

4. Чим визначається «схожість» об'єктів при вирішенні задачі кластеризації?

5. Що таке однорівнева і ієрархічна кластеризація?

6. Що таке чітка і нечітка кластеризація?

7. Які є підходи до розрахунку відстані між кластерами?

8. Що таке алгомеративна і дівізімна ієрархічна кластеризація?

9. Опишіть один з розглянутих методів, що вирішують завдання кластеризації?

10. Як оцінити якість побудованої моделі для завдання кластеризації?

**Зміст звіту**

1. Тема і мета роботи.

2. Завдання до роботи.

3. Результати виконання завдань.

4. Відповіді на контрольні запитання.

5. Висновки, що відображують результати виконання роботи та їх критичний аналіз.