МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет "Львівська політехніка"

Інтелектуальний аналіз даних за допомогою програмного пакета WEKA та MS Excel.

Алгоритми кластеризації k-внутрішніх середніх та ієрархічної кластеризації.

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторної роботи № 5

з курсу **"Системи інтелектуального аналізу та візуалізації даних"**

для студентів за освітньою програмою Комп'ютерні науки (Проектування і програмування інтелектуальних систем та пристроїв)

Затверджено на засіданні кафедри
"Системи автоматизованого проектування"

Протокол N 1 від 28.08.2023р.

ЛЬВІВ 2023

1. МЕТА РОБОТИ

Метою лабораторної роботи є вивчення та застосування трьох методів кластеризації в середовищі Weka: поділяючого методу кластеризації K-середніх (SimpleKMeans), ієрархічного методу кластеризації (HierarchicalClusterer) та ієрархічного методу кластеризації COBWEB (COBWEB). Студенти мають набути навичок роботи з цими алгоритмами, вміти прикладати їх для аналізу реальних даних та інтерпретувати отримані результати.

2. КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

2.1. Методи кластеризації, які розглядаються у лабораторній роботі.

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

- поділяючий метод кластеризації K-середніх (*SimpleKMeans*),
- ієрархічний метод кластеризації (*HierarchicalClusterer*),
- ієрархічний метод кластеризації COBWEB (*COBWEB*).

2.2 Поняття кластеризації.

Алгоритми кластеризації допомагають користувачеві відповісти на запитання наступного типу: «Який вік покупців, що віддають перевагу сріблястий BMW M5?». Необхідні дані можуть бути отримані при аналізі віку покупців і кольору обраних ними машин. Далі, на підставі отриманих даних, можна зробити висновок, що певна вікова група (наприклад, люди у віці від 22 до 30 років) віддає перевагу певному кольору BMW M5 (75% купують автомобілі синього кольору). Точно так, можна визначити, що інша вікова група (наприклад, люди у віці від 55 до 62 років) віддає перевагу сріблястому BMW (65% купують сріблясті автомобілі, а 20% купують сірі). Аналіз даних дозволить створити кластери за віковими групами і кольорами і визначити залежності між даними.

Кластеризація дозволяє розбити дані на групи, кожна з яких має певні ознаки. Метод кластерного аналізу використовується у тих випадках, коли необхідно виділити деякі правила, взаємозв'язки або тенденції у великих наборах даних. Залежно від потреб, ви можете виділити кілька різних груп даних. Однією з явних переваг кластеризації у порівнянні з класифікацією полягає у тому, що для розбиття множини на групи може використовуватися будь-який атрибут (метод класифікації використовує тільки певну підмножину атрибутів). В якості основного недоліку методу кластеризації слід зазначити той факт, що розробник моделі повинен заздалегідь вирішити, на скільки груп слід розбити дані. Для людини, яка не має уявлення про конкретний набір даних, прийняти таке рішення досить важко. Нам варто створити три групи або п'ять груп? А може, нам потрібно визначити десять груп? Може знадобитися кілька повторювань проб і помилок, для того щоб визначити оптимальну кількість кластерів.

Тим не менше, для середньостатистичного користувача кластеризація може виявитися найбільш корисним методом інтелектуального аналізу даних. Цей метод дозволить вам швидко розбити ваші дані на окремі групи і зробити конкретні висновки і припущення щодо кожної групи. Математичні методи, що реалізують кластерний аналіз, досить складні і заплутані, так що в разі кластеризації ми будемо цілком покладатися на обчислювальні можливості WEKA.

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

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.

2.4. Алгоритм кластеризації k-середніх (k-means)

K-середніх — це найпростіший алгоритм неконтрольованого навчання, який вирішує проблему кластеризації. Процедура полягає в простому та легкому способі класифікації заданого набору даних через певну кількість кластерів (припустимо k кластерів), фіксованих апріорі. Основна ідея полягає у визначенні k центрів, по одному для кожного кластера. Ці центри слід вибирати певним чином, оскільки різне розташування призводить до різного результату. Тому краще розмістити їх якомога далі один від одного. Наступним кроком буде взяти кожну точку, що належить до даного набору даних, і зв'язати її з найближчим центром. Коли вже не залишилося точок, тоді перший крок завершується, і ми отримуємо перший набір кластеризованих даних, який на наступних кроках буде уточнюватися.

- Алгоритм k-середніх — це алгоритм для кластеризації n об'єктів на основі атрибутів у k розділів, де k<n.
- Припускається, що атрибути об'єкта утворюють векторний простір.
- K – натуральне число.
- Групування здійснюється шляхом мінімізації суми квадратів відстаней між даними та відповідним центроїдом кластера.
- Кожен атрибут має бути приведений до нормального вигляду. Для цього кожен показник ділиться на різницю між найбільшим і найменшим значенням, які приймає розглянутий атрибут на конкретному наборі даних. Наприклад, якщо розглянутий атрибут - вік, і його найбільше значення - 72, а найменше - 16, то значенням 32 буде відповідати нормалізована величина 0.5714.

Алгоритм:

1. Вибирається кількість кластерів k.
2. З вихідної множини даних випадковим чином вибираються k спостережень, які служитимуть початковими центрами кластерів.
3. Для кожного спостереження вихідної множини визначається найближчий до нього центр кластера (відстань вимірюється в метриці Евкліда). У цьому записи, «притягнуті» певним центром, утворюють початкові кластери.
4. Обчислюються центроїди – центри тяжкості кластерів. Кожен центроїд — це вектор, елементи якого є середніми значеннями відповідних ознак, обчислені за всіма записами кластера.
5. Центр кластера зміщується до його центроїду, після чого центроїд стає центром нового кластера.
6. 3-й та 4-й кроки ітераційно повторюються. Очевидно, що на кожній ітерації відбувається зміна меж кластерів та усунення їх центрів. В результаті мінімізується відстань між елементами всередині кластерів та збільшуються міжкластерні відстані.

Зупинка алгоритму проводиться тоді, коли межі кластерів та розташування центроїдів не перестануть змінюватися від ітерації до ітерації, тобто, на кожній ітерації в кожному кластері

залишатиметься той самий набір спостережень. Насправді алгоритм зазвичай знаходить набір стабільних кластерів за декілька десятків ітерацій.

Перевагою алгоритму є швидкість та простота реалізації. До недоліків можна віднести невизначеність вибору початкових центрів кластерів, а також те, що число кластерів має бути задано спочатку, що може вимагати деякої апріорної інформації про вихідні дані.

Існують методи кластеризації, які можна розглядати як ті, що походять від k-середніх. Наприклад, у методі k-медіан (k-medians) для обчислення центроїдів використовується не середнє, а медіана, що робить алгоритм стійкішим до аномальних значень даних.

Алгоритм g-середніх (від gaussian) будує кластери, розподіл даних у яких прагне нормальному (гаусівському) і знімає невизначеність вибору початкових кластерів. Алгоритм C-середніх використовує елементи нечіткої логіки, враховуючи при обчисленні центроїдів як відстані, а й ступінь належності спостереження до безлічі об'єктів у кластері. Також відомий алгоритм Ллойда, який як початкове розбиття використовує не безліч векторів, а області векторного простору.

Ідея методу k-середніх була одночасно сформульована Гуго Штейнгаузом та Стюартом Ллойдом у 1957 р. Сам термін «k-середніх» був уперше введений Дж. Маккуїном у 1967 р.

2.4.1 Приклад 1. Голосування.

```
@attribute 'handicapped-infants' { 'n', 'y'}
@attribute 'water-project-cost-sharing' { 'n', 'y'}
@attribute 'adoption-of-the-budget-resolution' { 'n', 'y'}
@attribute 'physician-fee-freeze' { 'n', 'y'}
@attribute 'el-salvador-aid' { 'n', 'y'}
@attribute 'religious-groups-in-schools' { 'n', 'y'}
@attribute 'anti-satellite-test-ban' { 'n', 'y'}
@attribute 'aid-to-nicaraguan-contras' { 'n', 'y'}
@attribute 'mx-missile' { 'n', 'y'}
@attribute 'immigration' { 'n', 'y'}
@attribute 'synfuels-corporation-cutback' { 'n', 'y'}
@attribute 'education-spending' { 'n', 'y'}
@attribute 'superfund-right-to-sue' { 'n', 'y'}
@attribute 'crime' { 'n', 'y'}
@attribute 'duty-free-exports' { 'n', 'y'}
@attribute 'export-administration-act-south-africa' { 'n', 'y'}
@attribute 'Class' { 'democrat', 'republican'}
@data
'n','y','n','y','y','y','n','n','n','y',?,'y','y','y','n','y','republican'
```

1. Запустіть Weka Explorer і завантажте файл даних vote.arff в інтерфейс попередньої обробки.

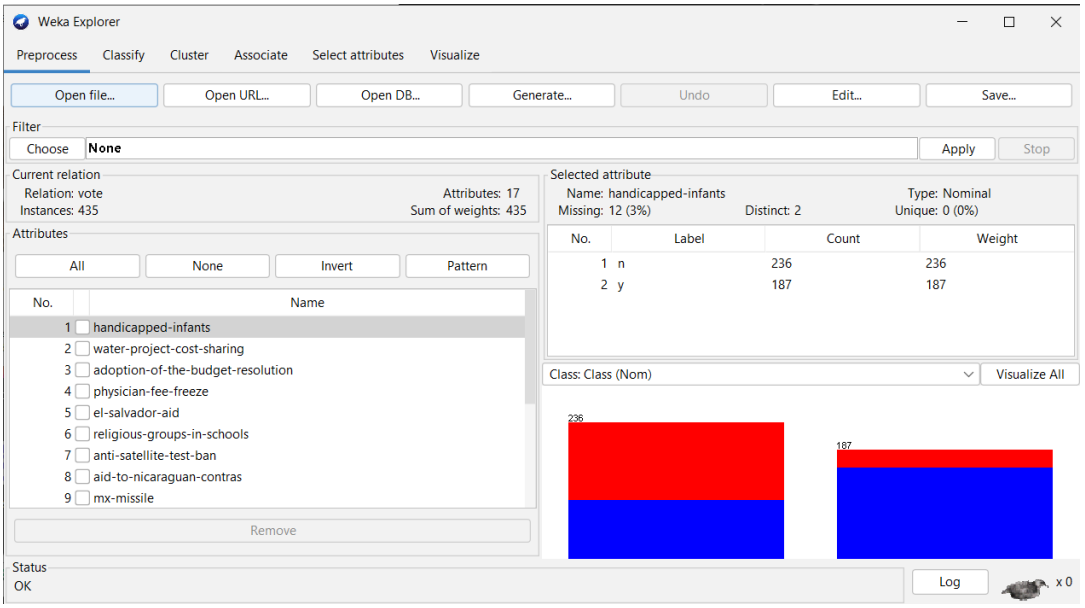

Рис. 1. Перегляд властивостей атрибуту

2. Щоб виконати кластеризацію, виберіть вкладку «кластер» у провіднику та натисніть кнопку вибору. Результатом цього кроку є спадний список доступних алгоритмів кластеризації.

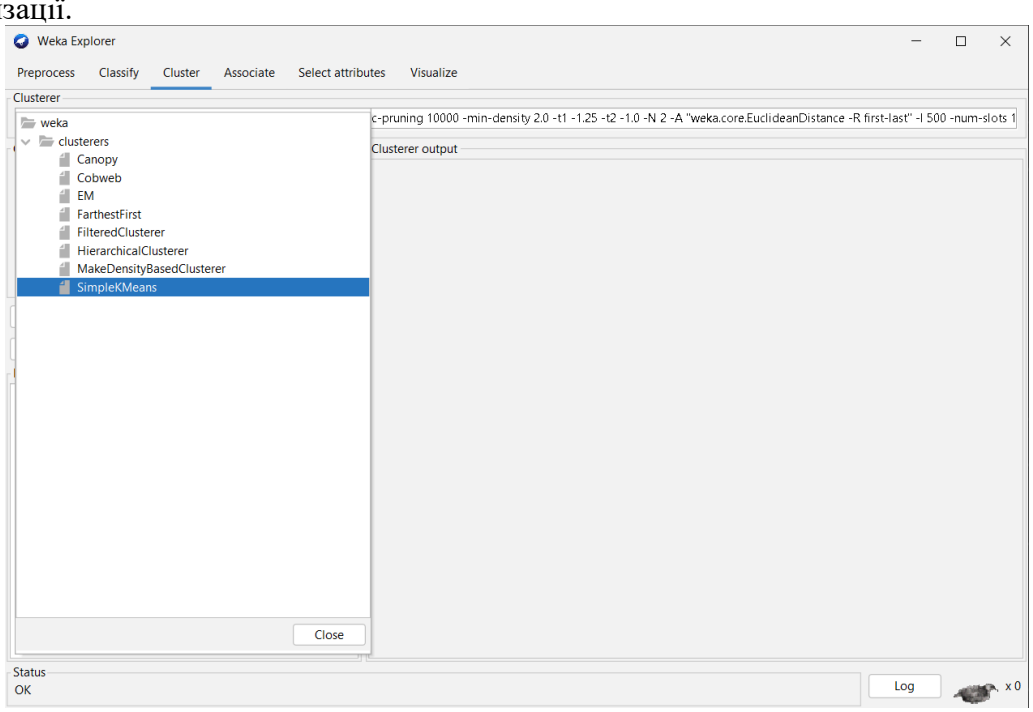

Рис.2. Вибір алгоритму кластеризації

3. У цьому випадку ми вибираємо «k-внутрішніх середніх».

4. Потім клацніть поле праворуч від кнопки вибору (Choose), щоб відкрити спливаюче вікно із набором параметрів алгоритму. У цьому вікні ми вводимо шість для кількості кластерів і залишаємо значення початкового числа без змін. Початкове значення використовується для генерації випадкового числа, яке використовується для внутрішніх призначень екземплярів кластерів.

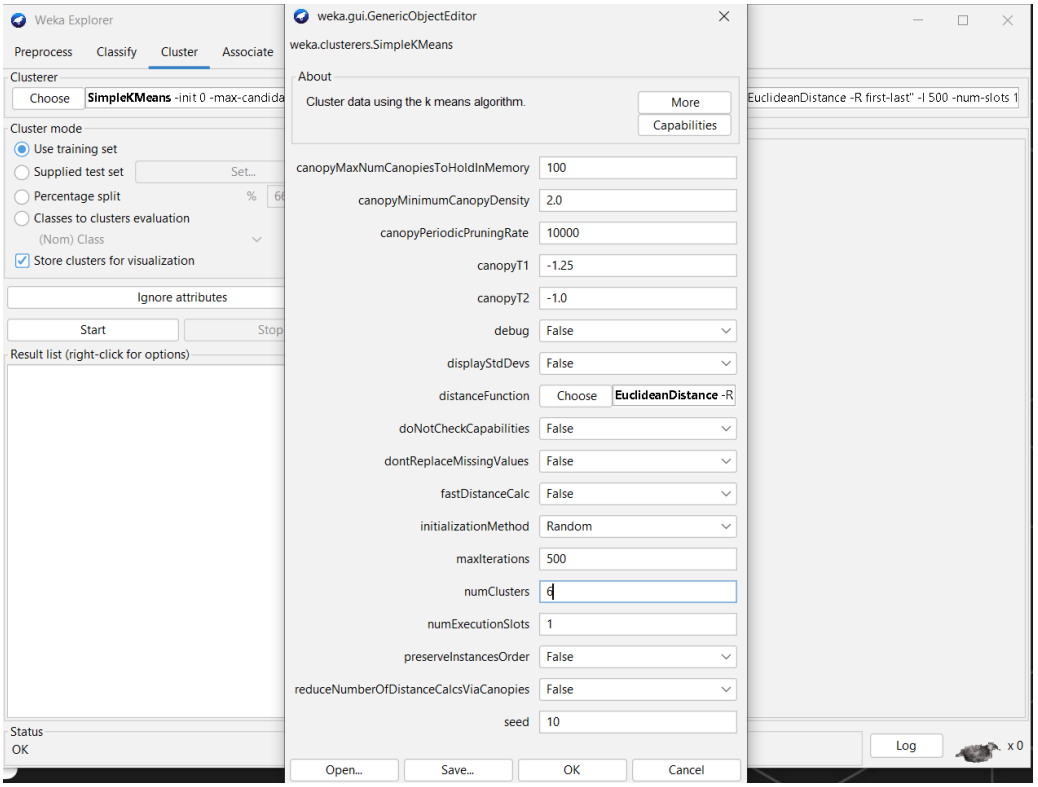

Рис. 3. Налаштування параметрів алгоритму

5. Один раз вказано параметр. Ми запускаємо там алгоритм кластеризації, ми повинні переконатися, що вони знаходяться на панелі «режим кластера». Вибирається параметр використання навчального набору, а потім ми натискаємо кнопку «Пуск». Цей процес і отримане вікно показано на наступних знімках екрана.

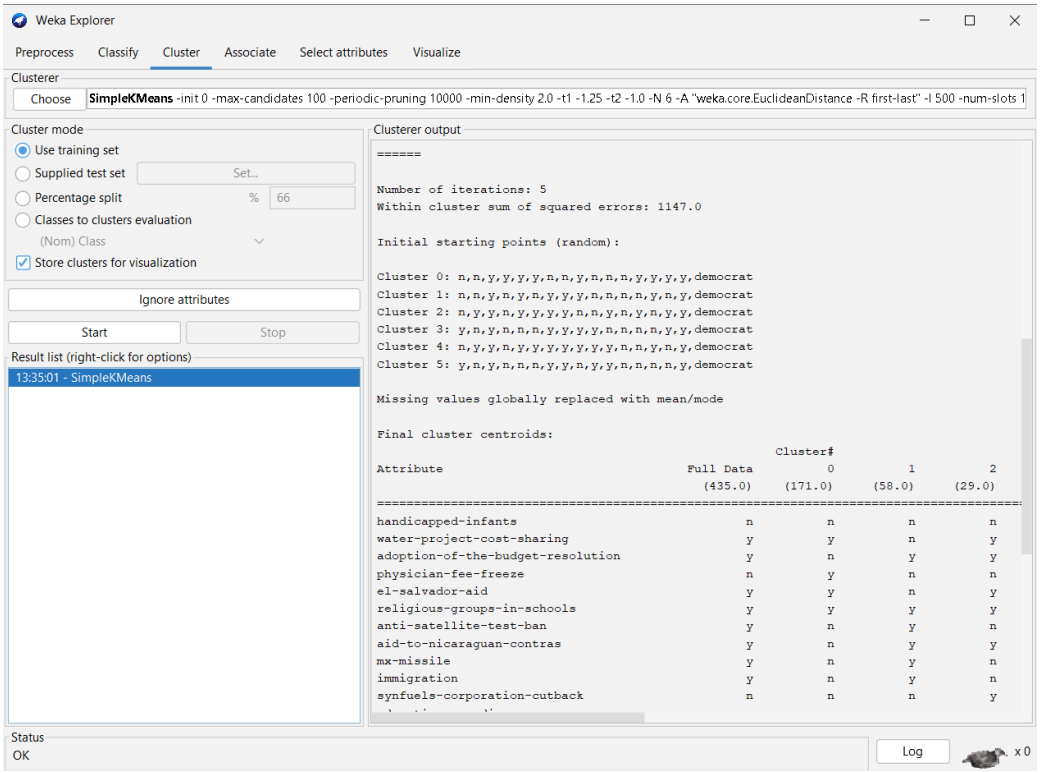

Рис. 4. Результати кластеризації.

Щоб візуалізувати отримані результати, виберіть Visualize і потім двічі клацніть по зоні візуалізації.

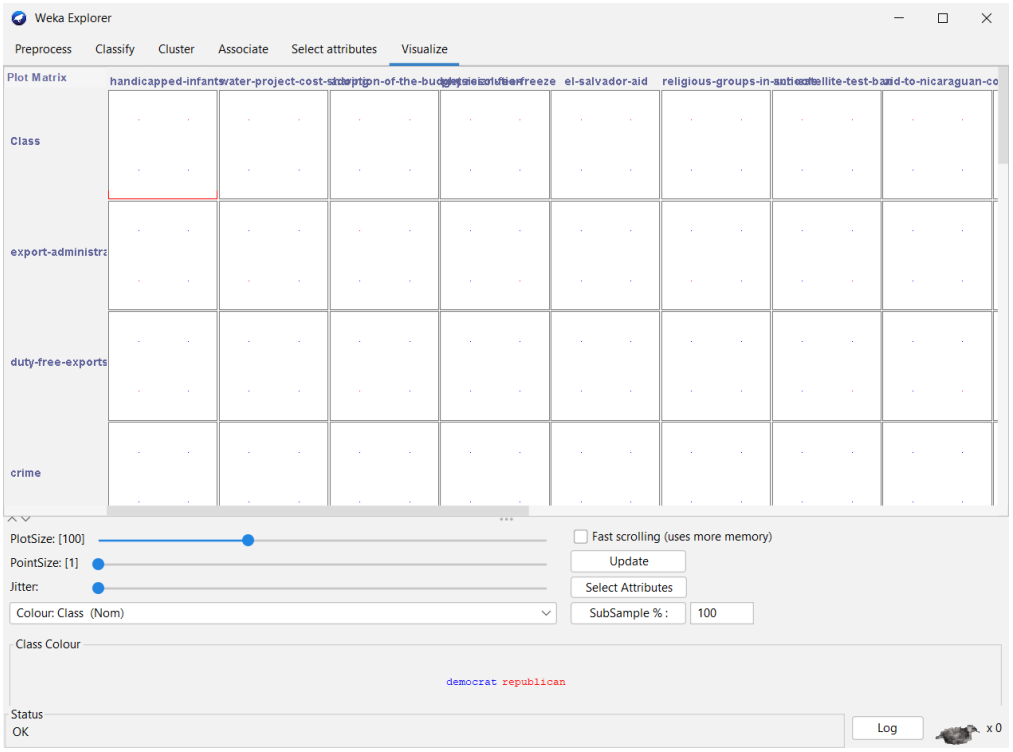

Рис.5. Графічне відображення результатів кластеризації

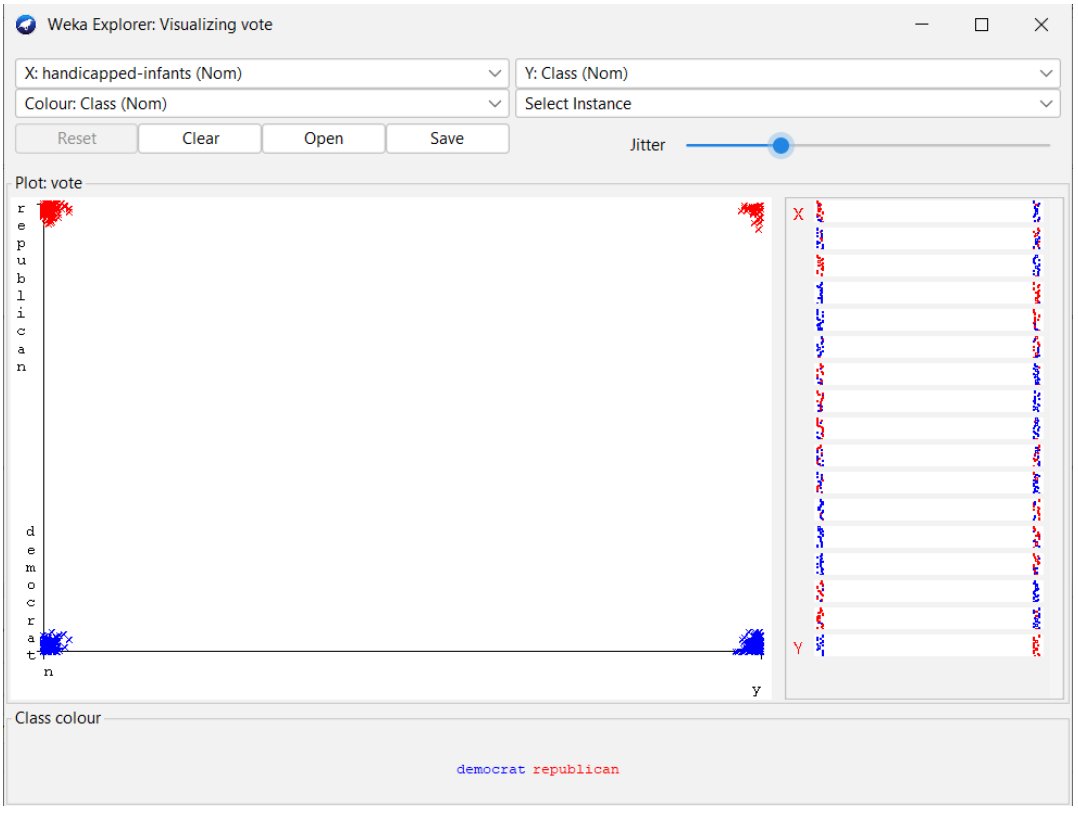

Рис. 6. Графічне відображення результатів кластеризації

У наведеному вище прикладі можна змінювати jitter, щоб переглядати результати кластеризації.

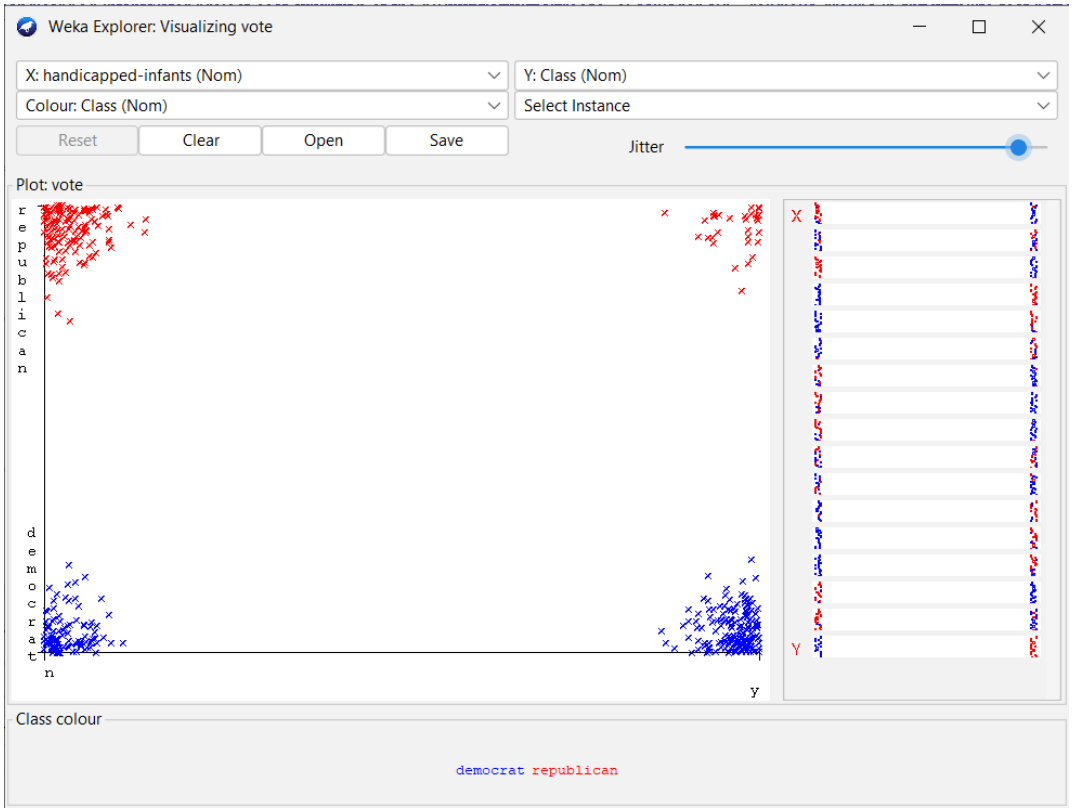

Рис. 7. Графічне відображення результатів кластеризації

2.4.2 Приклад 2. Дилерський центр BMW.

Для побудови моделі кластеризації скористаємося даними дилерського центру BMW. Співробітники центру зібрали дані про всіх відвідувачів демонстраційного залу, машинах, які їх зацікавили, і про те, наскільки часто відвідувачі демонстраційного залу у підсумку купували автомобіль, який їм сподобався. Тепер дилерському центру потрібно проаналізувати ці дані для того, щоб виділити різні групи відвідувачів і зрозуміти, чи не можна визначити будь-які тенденції в їх поведінці. У нашому прикладі використовується 100 записів, і кожен стовпець описує певний етап, який, як правило, проходить покупець у процесі вибору та придбання автомобіля. Відповідно, значення 1 у стовпці говорить про те, що відвідувач пройшов конкретний етап, а 0 – що відвідувач цей етап не пройшов. Файл з даними у форматі ARFF приведений нижче.

Дані для кластерного аналізу засобами WEKA

```
@Attribute Dealership numeric
@Attribute Showroom numeric
@Attribute ComputerSearch numeric
@Attribute M5 numeric
@Attribute 3Series numeric
@Attribute Z4 numeric
@Attribute Financing numeric
@Attribute Purchase numeric

@Data
1,0,0,0,0,0,0,0
1,1,1,0,0,0,1,0
…
```

Завантажте файл **bmw-browsers.arff** у WEKA. Після завантаження даних ваш екран WEKA повинен виглядати так, як показано на рис. 8.

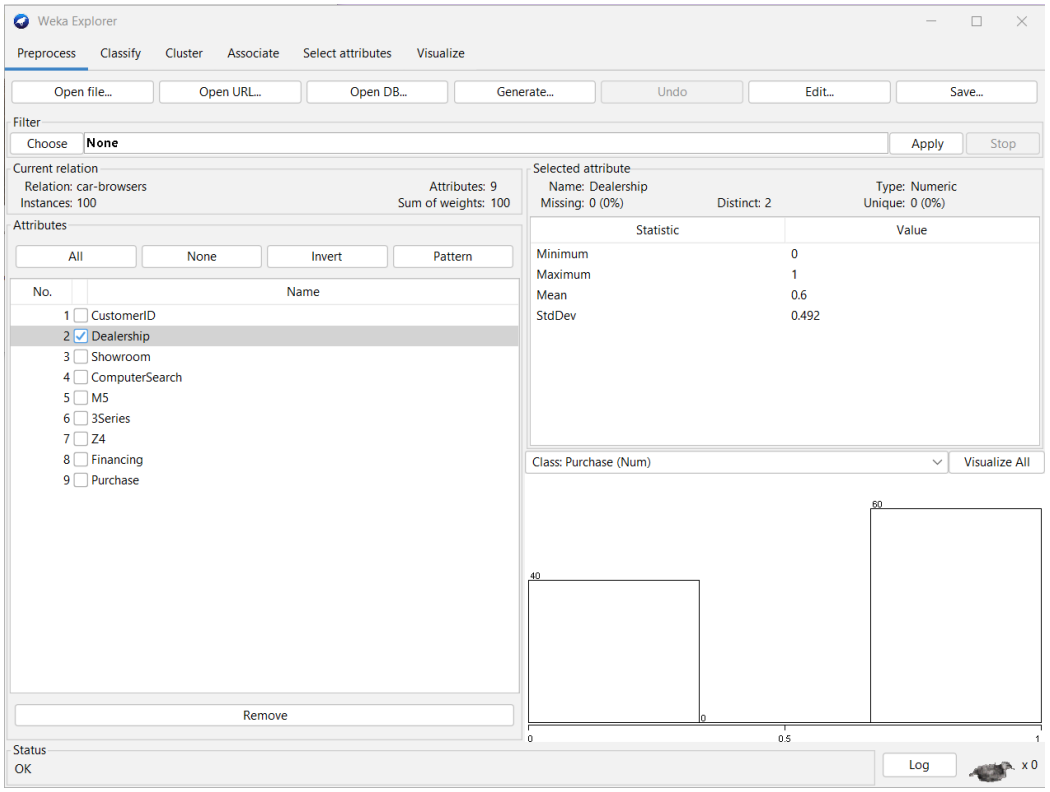

Рис. 8. Дані BMW для кластеризації

Оскільки ми хочемо розбити наявні у нас дані на кластери, замість закладки **Classify** нам буде потрібно закладка **Cluster**. Натисніть на кнопку **Choose** і в пропонованому меню виберіть опцію **SimpleKMeans** (в рамках даної роботи ми будемо користуватися цим методом кластеризації). У результаті вікно **WEKA Explorer** буде виглядати так, як показано на рис. 9.

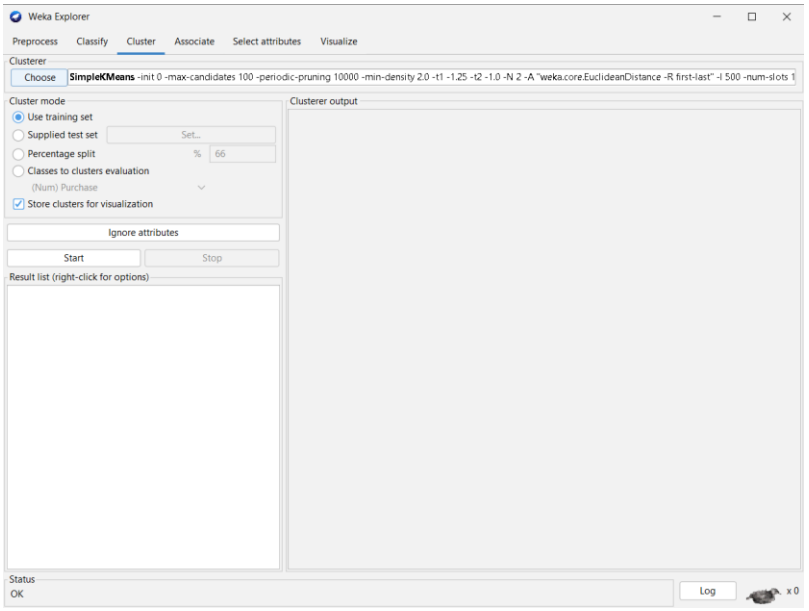

Рис. 9. Алгоритм кластеризації даних BMW

Тепер нам потрібно вибрати необхідні параметри для нашого алгоритму кластеризації. Клацніть на опції **SimpleKMeans**. Єдиний атрибут алгоритму, який нас цікавить – це поле **numClusters**, яке вказує на кількість кластерів для розбиття (нагадуємо, що це значення вам потрібно вибрати ще до створення моделі). Змінимо значення за замовчуванням (**2**) на **5**. Постарайтеся запам'ятати послідовність кроків, щоб ви змогли згодом змінити кількість кластерів. Тепер ваше вікно **WEKA Explorer** має виглядати так, як показано на рис. 10. Натисніть на кнопку **OK**, щоб зберегти вибрані параметри.

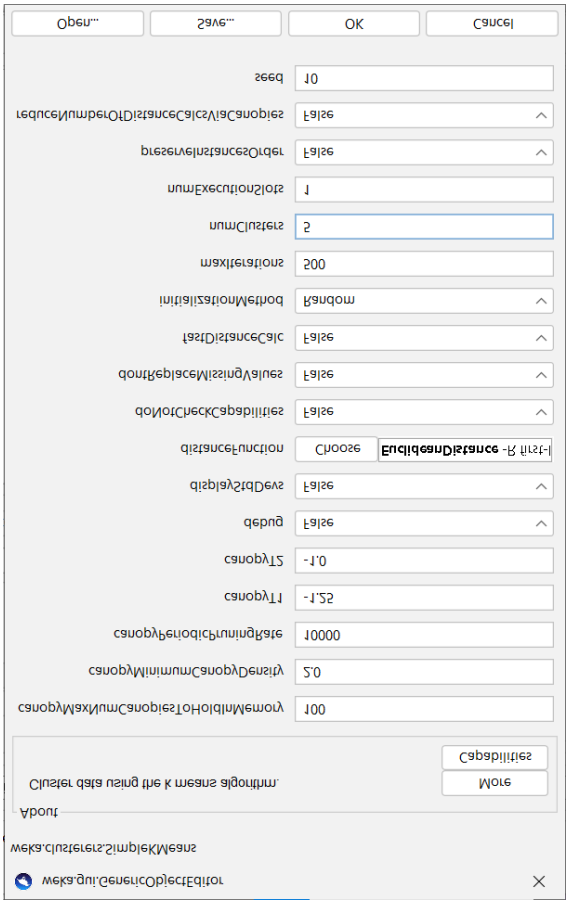

Рис. 10. Налаштування алгоритму кластеризації

Тепер ми можемо приступити до створення моделі. Як вже зазначалося вище, для розбиття 100 рядків на 5 кластерів за допомогою електронних таблиць було б потрібно декілька годин, проте WEKA видає нам результат менш ніж за секунду. Результат обробки наших даних показаний нижче.

Результати кластеризації

Clusterer output

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute         Full Data          0          1          2          3          4
                    (100.0)     (16.0)     (32.0)      (6.0)     (17.0)     (29.0)
=============================================================================
CustomerID             50.5     29.625      60.25    10.8333    26.8235    73.3448
Dealership              0.6     0.9375     0.6875          1          1          0
Showroom               0.72          1     0.7188     0.1667     0.1765          1
ComputerSearch         0.43     0.5625     0.2188          1     0.7059     0.3103
M5                     0.53      0.625     0.9688     0.8333     0.3529     0.0345
3Series                0.55      0.625     0.3125          1          0          1
Z4                     0.45     0.8125     0.1875        0.5     0.2353     0.6552
Financing              0.61        0.5     0.8125     0.6667     0.5294     0.4828
Purchase               0.39          0     0.6875        0.5     0.2941     0.3103



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       16 ( 16%)
1       32 ( 32%)
2        6 (  6%)
3       17 ( 17%)
4       29 ( 29%)
```

Як нам інтерпретувати отриманий результат? Дані кластеризації показують, яким чином сформований кожен кластер: значення «1» означає, що у всіх даних у цьому кластері відповідний атрибут дорівнює 1, а значення «0» означає, що у всіх даних у цьому кластері відповідний атрибут дорівнює 0. Дані відповідають середньому значенню атрибута у кластері. Кожен кластер характеризує певний тип поведінки клієнтів, таким чином, на підставі нашого розбиття ми можемо зробити деякі корисні висновки:

- **Кластер 0** - цю групу відвідувачів можна було б назвати «мрійники». Вони ходять навколо дилерського центру, розглядаючи машини, виставлені на зовнішній парковці, але ніколи не заходять всередину, і, гірше того, ніколи нічого не купують.
- **Кластер 1** - цю групу слід було б назвати «шанувальники M5», оскільки вони відразу ж підходять до виставлених автомобілів цієї моделі, повністю ігноруючи BMW серії 3 або Z4. Тим не менш, ця група не відрізняється високими показниками покупки машин - всього 52%. Це потенційно може свідчити про недостатньо продуману стратегію продажів і про необхідність покращити роботу дилерського центру, наприклад, за рахунок збільшення кількості продавців у секції M5.
- **Кластер 2** - ця група настільки мала, що ми могли б назвати її вибраковуванням. Справа в тому, що дані цієї групи статистично досить розкидані, і ми не можемо зробити будь-яких певних висновків щодо поведінки відвідувачів, що потрапили у цей кластер (подібна ситуація може вказувати на те, що вам слід скоротити кількість кластерів в моделі).
- **Кластер 3** - цю групу слід було б назвати «улюбленці BMW», тому що відвідувачі, що потрапили в цей кластер, завжди купують машину і отримують необхідне фінансування. Зверніть увагу, дані цього кластеру демонструють цікаву модель поведінки цих покупців: спочатку вони оглядають виставлені на парковці машини, а потім звертаються до пошукової системи дилерського центру. Як правило, вони купують моделі M5 або Z4, але ніколи не беруть моделі третьої серії. Дані цього кластеру вказують на те, що дилерському центру слід активніше привертати увагу до пошукових комп'ютерів (можливо, винести їх на зовнішню парковку), і крім того, слід знайти якийсь спосіб виділити моделі M5 і Z4 в результатах пошуку, щоб гарантовано звернути на них увагу відвідувачів. Після того, як відвідувач, що потрапив в цей кластер, вибрав певну модель автомобіля, він гарантовано отримує необхідний кредит і здійснює покупку.
- **Кластер 4** - цю групу можна назвати «початківці власники BMW», оскільки вони завжди шукають моделі 3 серії і ніколи не цікавляться більш дорогими M5. Вони відразу ж проходять в демонстраційний зал, не витрачаючи час на огляд машин на зовнішній стоянці. Крім того, вони не користуються пошуковою системою центру. Приблизно 50% цієї групи отримують схвалення по кредиту, тим не менш, покупку роблять всього 32% учасників. Аналізуючи дані цього кластеру, можна зробити наступний висновок: відвідувачі цієї групи хотіли б купити свій перший BMW і точно знають, яка машина їм потрібна (модель 3 серії з мінімальною конфігурацією). Однак, для того щоб купити машину, їм потрібно отримати позитивне рішення по кредиту. Щоб підвищити рівень продажів серед відвідувачів 4 кластера, дилерському центру слід було б знизити рівень вимог для отримання кредиту або знизити ціни на моделі 3 серії.

Ще один цікавий спосіб вивчення результатів кластеризації - це візуальне подання даних. Клацніть правою кнопкою мишки в секції **Result List** закладки **Cluster**. У контекстному меню виберіть опцію **Visualize Cluster Assignments**. У результаті відкриється вікно з графічним представленням результатів кластеризації, налаштування якого ви можете вибрати найбільш зручним для вас чином. Для нашого прикладу, змініть налаштування осі **X** так, щоб вона відповідала кількості автомобілів M5 (**M5 (Num)**), а налаштування осі **Y** - так, щоб вона показувала кількість куплених автомобілів (**Purchase (Num)**), і вкажіть виділення кожного кластера окремим кольором (для цього встановіть значення поля **Color** в **Cluster (Nom)**). Такі налаштування допоможуть нам оцінити розподіл по кластерах залежно від того, скільки людина цікавилося BMW M5, і скільки людей купило цю модель. Крім того, посуньте покажчик **Jitter** приблизно на три чверті у бік максимуму, це штучно збільшить розкид між групами точок, щоб нам було зручніше їх переглядати.

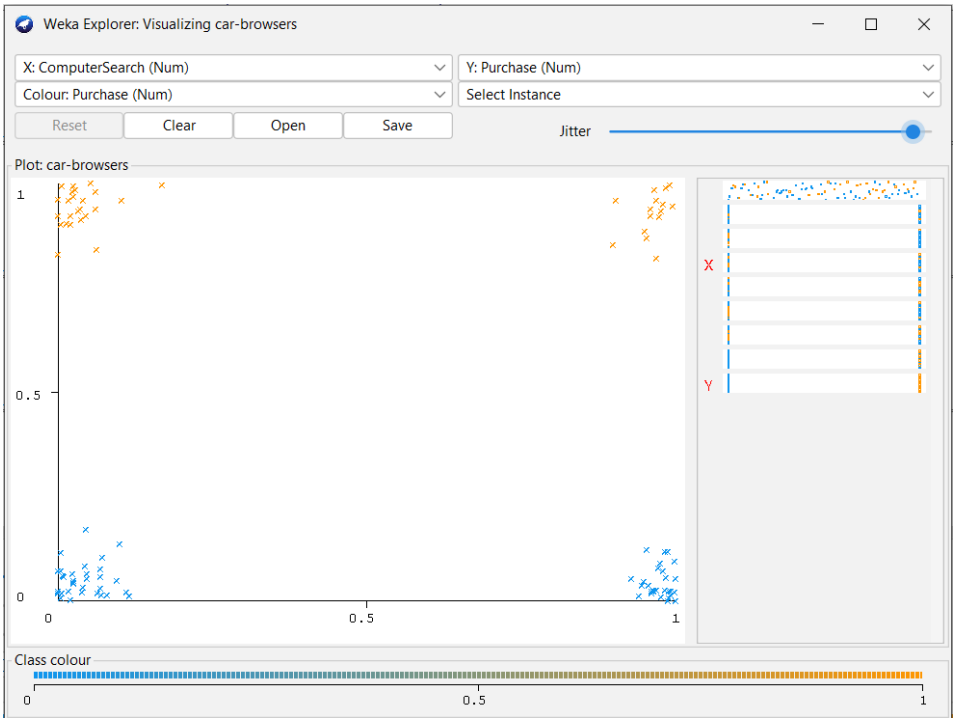

Рис. 11. Візуальне представлення кластеризації

Чи відповідає візуальне відображення кластеризації тим висновкам, які ми зробили на підставі даних в одержаному результаті кластеризації? Як ми бачимо, поблизу точки **X = 1**, **Y = 1** (відвідувачі, які цікавилися автомобілями моделі M5 і купили їх) розташовані тільки два кластери: **1** і **3**. Аналогічно, поблизу точки **X = 0**, **Y = 0** розташовані тільки два кластери: **4** і **0**. Чи відповідає це нашим висновкам? Так, відповідає. Кластери **1** і **3** купують BMW M5, у той час як кластер **0** не купує нічого, а кластер **4** шукає BMW серії 3. На рис. 4 показано візуальне відображення кластерів нашої моделі. Пропонуємо вам самостійно попрактикуватися у виявленні інших трендів і течій, змінюючи налаштування осей **X** і **Y**.

2.5. Ієрархічна кластеризація.

Цей метод кластеризації дає відповідь на питання наступного типу: «Яка ймовірність того, що хтось купить останню модель BMW M5?» Створення класифікаційної моделі (дерева рішень) дозволяє оцінити ймовірність того, що якась людина купить автомобіль моделі M5. В якості вузлів дерева можуть використовуватися такі показники, як вік, рівень доходу, кількість вже наявних машин, сімейний стан, діти, наявність власного будинку або оренда будинку. Параметри конкретної людини будуть використовуватися для проходження по дереву з метою визначення ймовірності покупки M5.e

Ієрархічні алгоритми кластеризації, або алгоритми таксономії, будують не одне розбиття вибірки на непересічні класи, а систему вкладених розбиттів. Результат таксономії зазвичай представляється у вигляді таксономічного дерева — дендрограми. Класичним прикладом такого дерева є ієрархічна класифікація тварин і рослин. Дендограми дозволяє уявити кластерну структуру у вигляді плаского графіка незалежно від того, яка розмірність початкового простору. Існують і інші способи візуалізації багатовимірних даних, такі як багатовимірне шкалювання або карти Кохонена, але вони привносять в картину штучні спотворення, вплив яких досить важко оцінити. Є два типи методів:

1. Агломеративні методи: нові кластери утворюються шляхом об'єднання дрібніших кластерів, і таким чином дерево створюється від листя до стовбура.

2. Дивізійні методи: нові кластери створюються шляхом ділення більших кластерів на більш дрібні, і таким чином дерево створюється від стовбура до листя.

Подібність кластерів часто розраховується через «неподібність», наприклад, евклідова відстань між двома кластерами. Отже, чим більше відстань між двома кластерами, тим краще. Ключовою операцією в ієрархічній агломераційній кластеризації є неодноразове об'єднання

двох найближчих кластерів у один кластер, але дуже важливо спочатку відповісти на три питання: як ви представити кластер з більш ніж однією точкою, як визначити «близькість» кластерів та коли перестати поєднувати кластери. Злиття кластерів припиняється в залежності від доступної інформації про дані, які ми маємо. Якщо групувати футболістів на полі на основі їхніх позицій на полі, яке представлятиме їх координати для розрахунку відстані між гравцями, очевидно, що треба зупинитися на лише двох кластерах, оскільки можуть бути тільки дві команди, які грають у футбольний матч. Алгоритм методу виглядає таким чином:

1. Обчислити матрицю близькості, що містить відстань між кожною парою шаблонів. Розглядати кожен зразок як окремий кластер.

2. Знайти найбільш схожу пару кластерів за допомогою матриці близькості. Об'єднати ці два кластера в один більший кластер. Оновити матрицю близькості, щоб відобразити цю операцію злиття.

3. Якщо всі шаблони знаходяться в одному кластері, зупинитися. В іншому випадку перейти до кроку 2. Ієрархічний алгоритм дає дендограму, що представляє собою складене групування шаблонів і рівні схожості, на яких змінюються самі групування. Більшість ієрархічних алгоритмів кластеризації є варіантами однозв'язного і повнозв'язного алгоритму, а також алгоритму мінімальної дисперсії. Найбільш популярними з них є однозв'язний та повнозв'язний алгоритми. Вони відрізняються тим, як вони характеризують подібність між парою кластерів.

2.5.1 Приклад 3. Голосування.

1. Відкрийте файл даних у Weka Explorer. Передбачається, що необхідні поля даних були дискретизовані. Клацнувши вкладку кластера, ви відкриєте інтерфейс для алгоритму кластера.

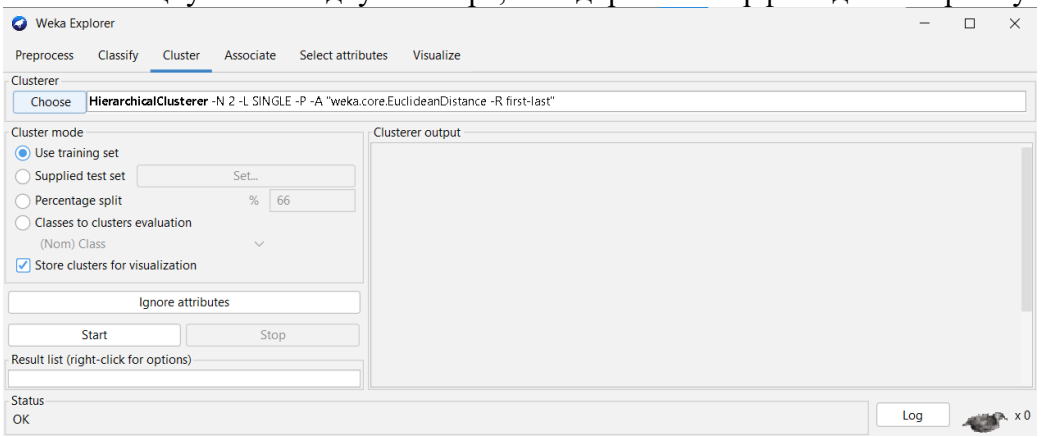

Рис. 12 Вибір ієрархічного алгоритму кластеризації

2. Ми будемо використовувати ієрархічний алгоритм кластеризації. Аналогічно попереднім прикладам налаштуйте параметри кластеризації. Встановіть кількість кластерів = 5.

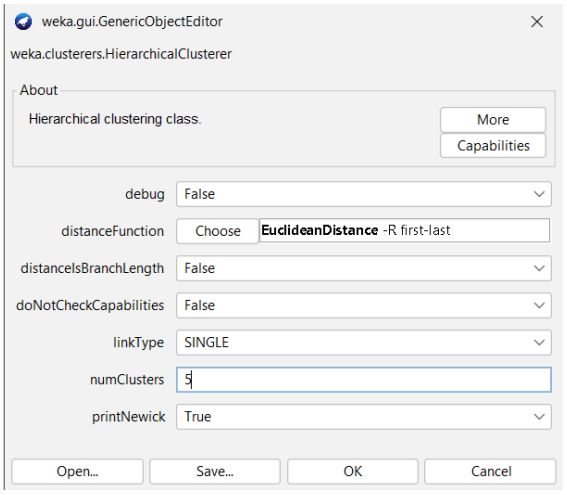

Рис.13 Налаштування параметрів кластеризації

4. Запустіть кластеризацію та візуалізуйте отримані результати шляхом вибору Visualize tree.

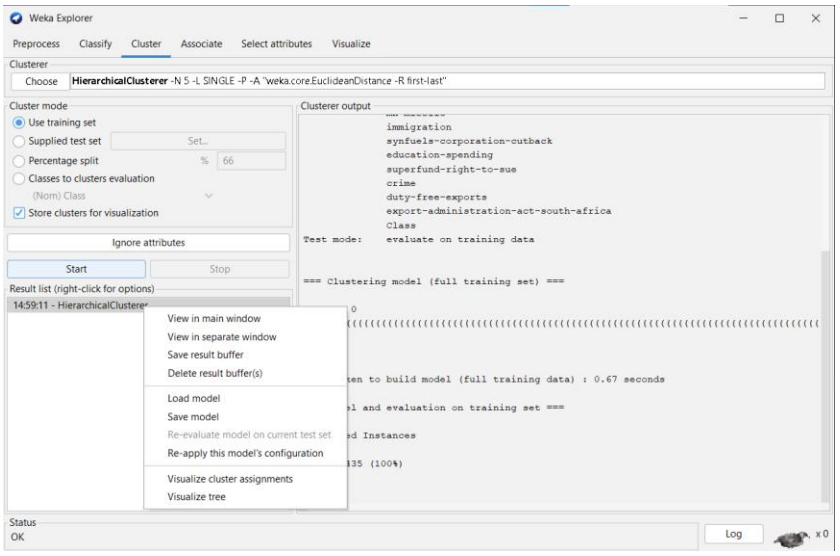

Рис. 14 Вибір Visualize tree

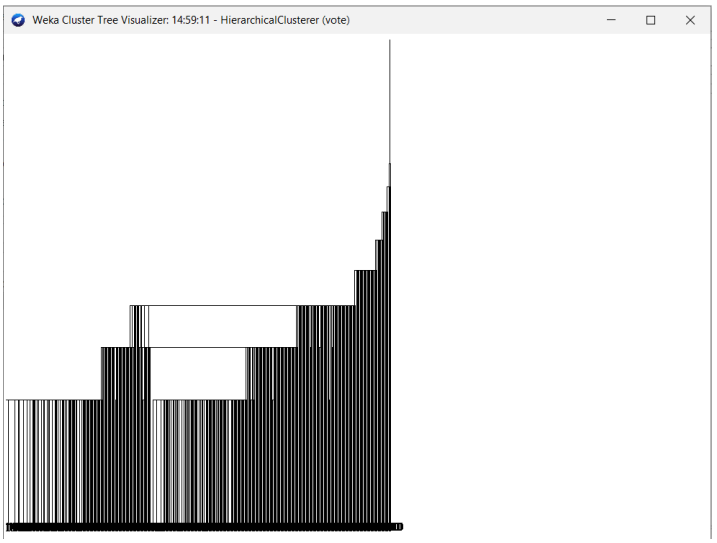

Рис. 15 Побудоване дерево кластеризації

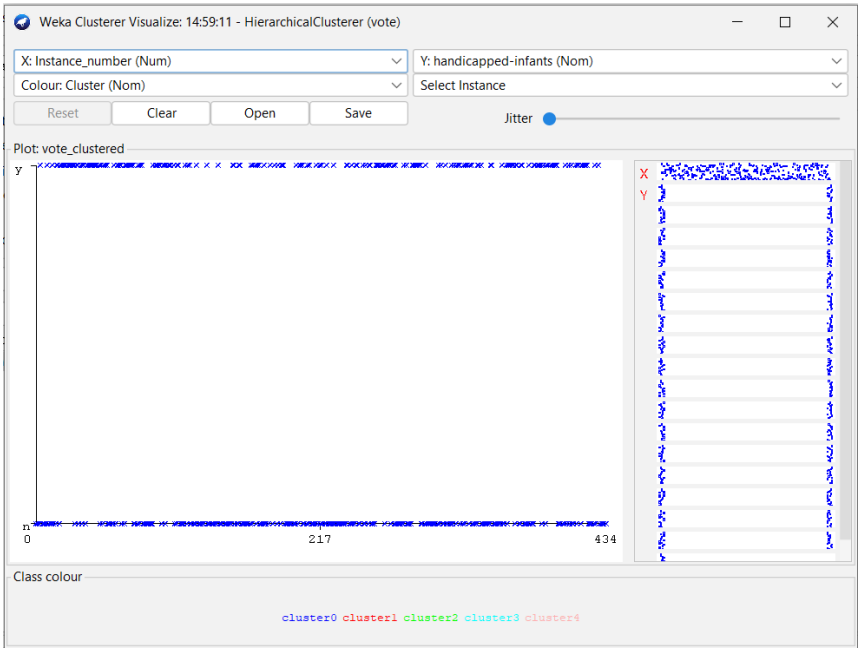

Рис. 16 Вибір Visualize cluster assignments

2.6. Параметри налаштування алгоритмів класифікації

Розглянемо параметри налаштування використовуваних алгоритмів кластеризації в WEKA (табл.1).

Таблиця 1. Параметри налаштування кластеризаторів

Метод	**Параметр**
SimpleKMeans	displayStdDevs – відображати значення стандартного відхилення для числових атрибутів і підрахунки для номінальних атрибутів. distanceFunction – функція відстані. dontReplaceMissingValues – не замінювати пропущені значення середнім значенням або модою. maxIterations – максимальна кількість ітерацій алгоритму. numClusters – кількість кластерів. preserveInstancesOrder – зберігати порядок примірників у вибірці. seed – випадковий сід для рандомізації вибірки.
Hierarchical Clusterer	*distanceFunction* – функція відстані. *distanceIsBranchLength* – у дендрограмі висота лінії, що зв'язує кластери, буде показувати відстань між ними. *linkType* – тип зв'язку для розрахунку відстані між двома кластерами. *numClusters* – кількість кластерів. *printNewick* – виводити кластери у форматі Newick.
COBWEB	*acuity* – мінімальне значення стандартного відхилення для числових атрибутів. *cutoff* - встановити поріг до якого відсікати вузли дерева. *saveInstanceData* – зберегти інформацію про примірники для візуалізації. *seed* – випадковий сід для рандомізації вибірки.

3. ЛАБОРАТОРНЕ ТРЕНУВАННЯ

Перед тим, як розпочати виконання свого індивідуального завдання, проробіть вправи із наступними тестовими даними:

Вправа 1

1. Виконайте наступні завдання для набору даних ‘bank.arff’:

а. Запустіть алгоритм кластеризації SimpleKMeans, задаючи значення параметра K (кількість кластерів) від 1 до 12.

б. Запишіть в таблицю значення сум квадратичних помилок, одержуваних при різних значеннях K. Що означає цей параметр і як змінюються його значення?

в. Для значення K=5 вкажіть:

- 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.

г. 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 неодружені (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?

Вправа 2

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

а. 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.

Вправа 3. Ієрархічна кластеризація

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

- 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;
- вкажіть, що спільного у прапорів, що опинилися в одному кластері.

2. 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.

4. ЛАБОРАТОРНЕ ЗАВДАННЯ

Ваше завдання для цієї лабораторної роботи - оцінити алгоритми кластеризації за допомогою Weka.

1. Використайте набір даних, який ви вибрали для лабораторної №1. Якщо він не підходить для задач кластеризації, то виберіть інший, який підходить.

2. Визначте, як ви будете вимірювати якість сформованих кластерів.

3. Для свого набору даних застосуйте три алгоритми кластеризації та порівняйте їх результати, використовуючи ваші показники якості.

4. Напишіть короткий звіт:

a. Опишіть набори даних та ваші показники якості.

b. Опишіть налаштування експерименту, наприклад, як ви попередньо обробили дані (якщо такі є), як вибрали параметри для вибраних алгоритмів (якщо такі є) та чому.

c. Представити результати експерименту. Вони не повинні бути простим копіюванням та вставкою з вихідних даних Weka, а скоріше представленими у вигляді таблиці або діаграми для зручності порівняння.

d. Запропонуйте ідеї та зробіть висновки зі своїх експериментів. Наприклад, чи різні методи кластеризації мають різницю щодо якості або продуктивності для певних наборів даних, які ви вибрали? І чому? Як може допомогти попередня обробка даних? Чи існують умови або загальні типи наборів даних, які роблять певні алгоритми більш придатними, ніж інші?

Бібліотеки наборів даних:

1. UCI Machine Learning Repository: https://archive.ics.uci.edu/
2. Datasets section at Weka: http://www.cs.waikato.ac.nz/ml/weka/

5. КОНТРОЛЬНІ ЗАПИТАННЯ

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 дівізимна ієрархічна кластеризація?
9. Опишіть один з розглянутих методів, що вирішують завдання кластеризації?
10. Як оцінити якість побудованої моделі для завдання кластеризації?

6. ЗМІСТ ЗВІТУ

1. Тема і мета роботи.
2. Завдання до роботи.
3. Результати виконання завдань.
4. Висновки, що відображують результати виконання роботи та їх критичний аналіз.

7. КОРИСНІ РЕСУРСИ

1. https://www.youtube.com/watch?v=vYUDYLBIvBE
2. https://www.tutorialspoint.com/weka/weka_clustering.htm
3. http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html
4. https://www.softwaretestinghelp.com/weka-explorer-tutorial/
5. http://modelai.gettysburg.edu/2016/kmeans/assets/iris/Clustering_Iris_Data_with_Weka.pdf
6. https://storm.cis.fordham.edu/~yli/documents/CISC4631Spring16/Weka_LabFour.pdf
7. https://cs.ccsu.edu/~markov/ccsu_courses/DataMining-Ex3.html
8. http://syllabus.cs.manchester.ac.uk/ugt/2020/COMP33111/tutorials/Tutorial-7-Clustering.pdf
9. https://www.slideshare.net/ishanawadhesh/classification-and-clustering-analysis-using-weka
10. https://edumine.wordpress.com/2014/08/10/clustering-with-weka-3-6-part-1/