Лабораторна робота №3

Знайомство з методами кластеризації даних

Мета роботи: Ознайомитися та отримати навички кластеризації даних за допомогою Data Mining GUI бібліотеки WEKA.

Завдання: Виконати кластеризацію тестових даних за допомогою методу кластеризації.

Загальні відомості

Метод кластеризації - це метод аналізу даних, який дозволяє розділити екземпляри даних за значеннями їх атрибутів на класи, кожен з яких має певні ознаки. Кластерний аналіз використовується в тих випадках, коли необхідно автоматично виділити деякі правила, взаємозв'язки або тенденції у сукупності даних. В якості моделі кластеризації можна використовувати двовимірний простір Евкліда, в якому відносне скупчення екземплярів даних являє собою певний клас (див. рис. 1).

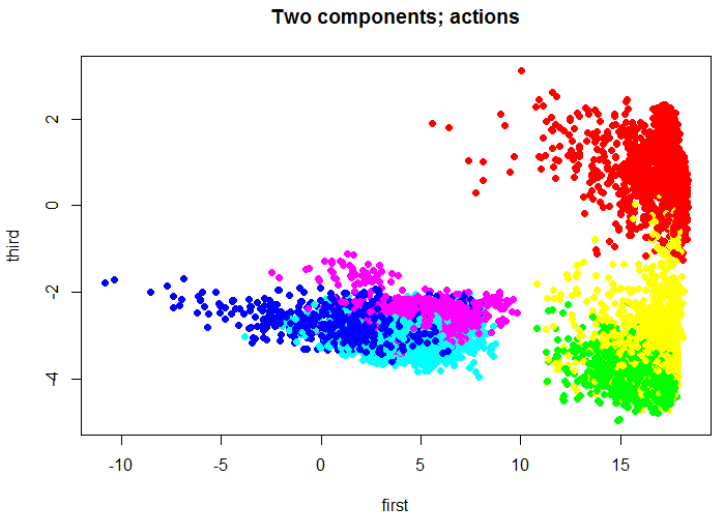

Рис. 1. Приклад результату кластеризації екземплярів даних

Просторовий підхід використовує *некласифіковані* екземпляри для аналізу зв'язків між значеннями їх атрибутів. Побудовану модель кластеризації можна використовувати для оцінювання приналежності нових екземплярів даних до вже відомих кластерів.

Набір даних для WEKA

Набір даних, який буде використаний для виконання кластерного аналізу, містить інформацію про усіх відвідувачів демонстраційного залу, про машини, які їх зацікавили, про те, наскільки часто відвідувачі купували цікаві їм автомобілі. Тепер дилерському центру треба проаналізувати ці дані для того, щоб виділити різні групи відвідувачів і зрозуміти, чи можна визначити деякі тенденції в їх поведінці.

У демонстраційному прикладі використовується 100 екземплярів, і кожен стовпець описує певний крок, який, як правило, проходить покупець у процесі вибору та придбання автомобіля. Відповідно, значення 1 та 0 показують чи відвідувач виконав конкретну дію, чи ні. Частина опису тестових даних у форматі ARFF зображена на ліст. 1.

```
@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
...
```

Лістинг 1. Файл даних для кластеризації у WEKA

Тестові дані у форматі ARFF можна знайти за адресою:
http://repository.seasr.org/Datasets/UCI/arff/

При виконанні кластеризації кожен атрибут має бути приведений до нормального вигляду. Для цього кожен показник ділиться на різницю між найбільшим і найменшим значенням, які приймає розглянутий атрибут для конкретного набору даних. Наприклад, якщо розглянутий атрибут — кількість років, відповідно при найбільшому значені - 72, а найменшому - 16, значенню атрибуту 32 буде відповідати нормалізована величина 0.5714.

Завантаження даних у WEKA

Коли файл з навчальними даними готовий, його потрібно завантажити у WEKA. Запустіть WEKA і виберіть опцію «Explorer». В результаті відкриється закладка «Preprocess» вікна «Explorer». Натисніть на кнопці Open File і виберіть створений вами ARFF-файл. Вікно WEKA «Explorer» з завантаженими даними показано на рис. 2.

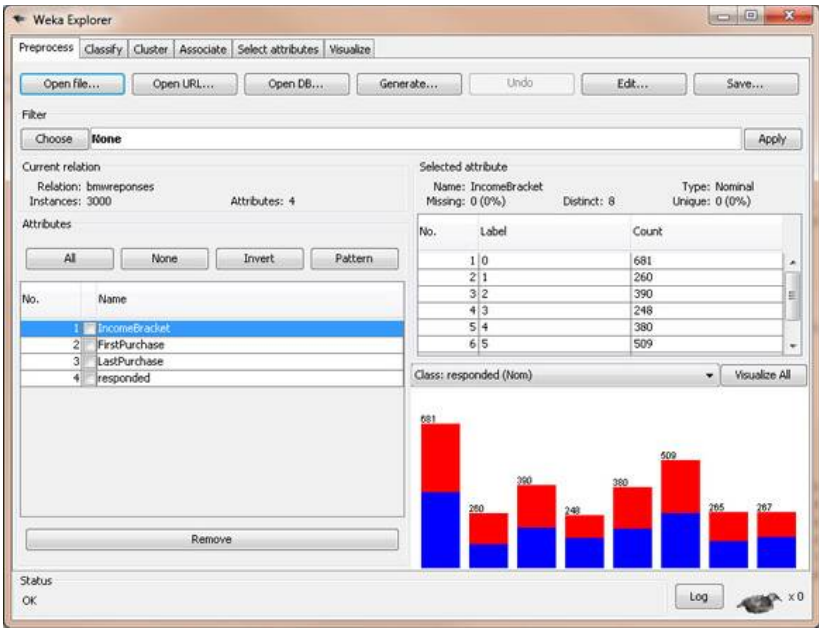

Рис. 2. Дані дилерського центру BMW для кластеризації у WEKA

У цьому вікні ви можете перевірити дані, на підставі яких ви збираєтеся будувати модель. У лівій частині вікна «Explorer» показані параметри об'єктів (Attributes), які відповідають заголовкам стовпців вихідної таблиці, а також вказано кількість об'єктів (Instances), тобто рядків таблиці. Якщо виділити мишкою один з заголовків стовпців, то в правій панелі буде виведена повна інформація про набір даних в даному стовпці. Також є можливість візуального аналізу даних (Visualize All).

Кластеризація у WEKA

Виберіть метод кластеризації: відкрийте закладку Cluster, в меню виберіть опцію відповідно до варіанту (див. табл. 1). В результаті вікно WEKA Explorer буде виглядати так, як зображено на рис. 3.

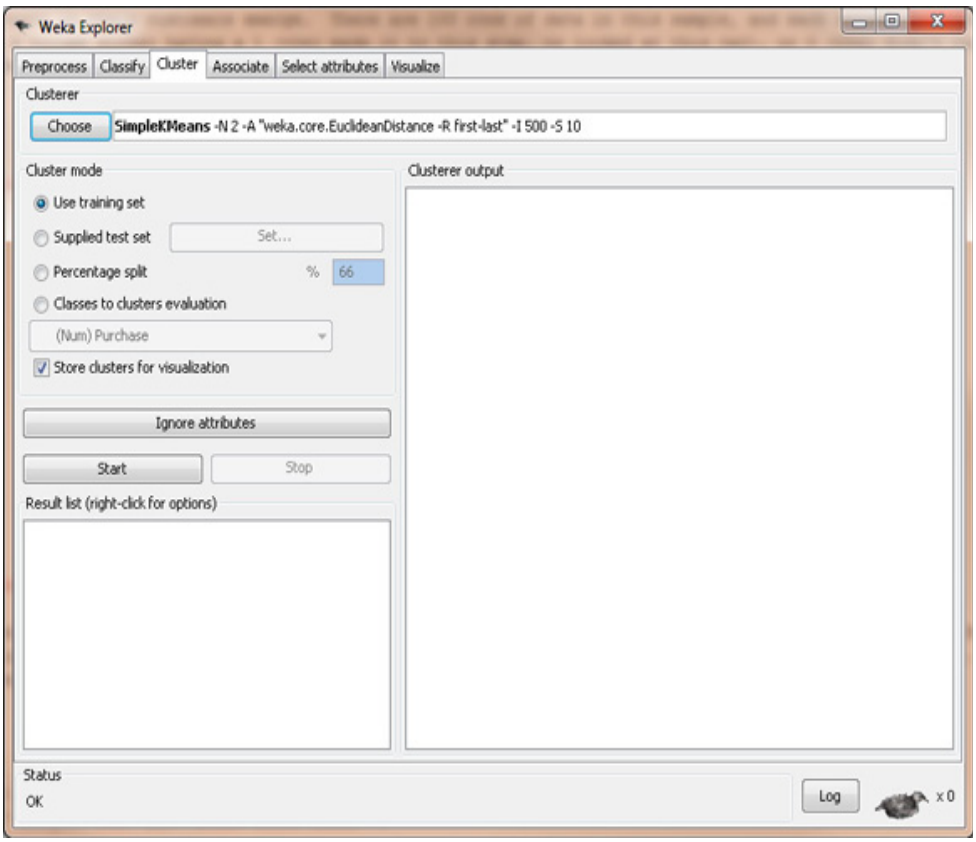

Рис. 3. Вікно вибору методу кластеризації даних

Тепер можна розпочати виконання кластерного аналізу засобами пакету WEKA. Переконайтеся, що обрана опція «Use training set», для того щоб пакет

WEKA при створенні моделі використовував саме ті дані, які ми тільки що завантажили у вигляді файлу.

Тепер потрібно вибрати необхідні параметри методу кластеризації. Натисніть на назві методу та вкажіть значення в полі numClusters, яке визначає кількість кластерів для розділення (нагадуємо, що це значення потрібно вибрати ще до створення моделі). Натисніть кнопку OK, щоб зберегти вибрані параметри.

Натисніть кнопку «Start». Результуюча модель повинна виглядати так, як показано на ліст. 2.

```
                                        Cluster#
Attribute       Full Data       0       1       2       3       4
                  (100)        (26)    (27)    (5)     (14)    (28)
=============================================================================
Dealership          0.6       0.9615  0.6667    1     0.8571    0
Showroom           0.72       0.6923  0.6667    0     0.5714    1
ComputerSearch     0.43       0.6538    0       1     0.8571  0.3214
M5                 0.53       0.4615  0.963     1     0.7143    0
3Series            0.55       0.3846  0.4444  0.8     0.0714    1
Z4                 0.45       0.5385    0     0.8     0.5714  0.6786
Financing          0.61       0.4615  0.6296  0.8       1     0.5
Purchase           0.39         0     0.5185  0.4       1     0.3214

Clustered Instances

0      26 ( 26%)
1      27 ( 27%)
2       5 (  5%)
3      14 ( 14%)
4      28 ( 28%)
```

Лістинг 2. Результат виконання кластеризації у WEKA

Один зі способів аналізу результатів кластеризації - це візуальне подання даних (див. рис. 4). Натисніть правою кнопкою миші в секції Result List закладки Cluster. У контекстному меню виберіть опцію Visualize Cluster Assignments. В результаті відкриється вікно з графічним представленням результатів кластеризації, налаштування ви можете вибрати найбільш зручним для вас чином. Для нашого прикладу, встановіть за віссю X атрибут кількості автомобілів M5 (Num), а за віссю Y атрибут кількості куплених автомобілів Purchase (Num) і

вкажіть виділення кожного кластеру окремим кольором, для цього встановіть значення поля Color в Cluster (Nom). Така послідовність дій допоможе оцінити розподіл по кластерах залежно від того, скільки людей цікавилося BMW M5, і скільки купило цю модель. Посуньте покажчик Jitter приблизно на три чверті в бік максимуму, це штучним чином збільшить розкид між групами точок, щоб було зручніше їх переглядати.

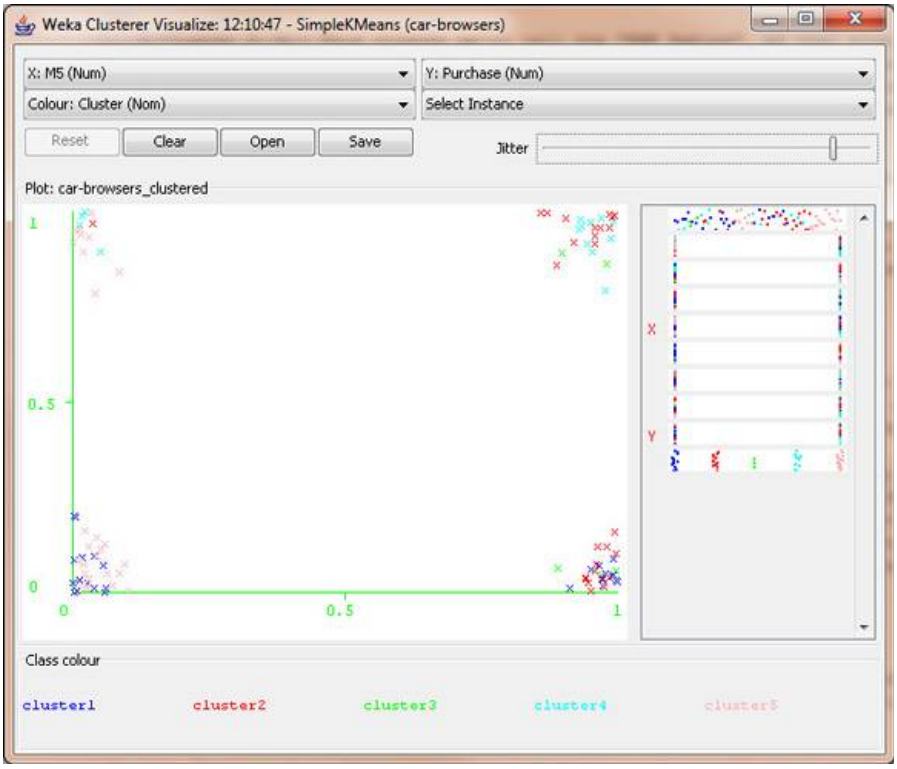

Рис. 4. Графічне відображення результату кластеризації

Метод кластеризації використовує відомі дані для аналізу зв'язків значень атрибутів. Коли з'являється новий екземпляр даних, потрібно лише оцінити дані за допомогою побудованої раніше моделі кластеризації та визначити приналежність даних до певного кластеру.

Таблиця 1

Варіанти завдання

Номер	Метод
1	SimpleKMeans
2	FarthestFirst
3	Hierarchical
4	FarthestFirst
5	SimpleKMeans
6	Hierarchical
7	SimpleKMeans
8	J48Hierarchical
9	FarthestFirst
10	SimpleKMeans

Зміст протоколу виконаної роботи:

1. титульний аркуш;
2. мета роботи;
3. інформація про дані та атрибути екземплярів даних;
4. скриншот статистики за даними;
5. скриншот опису побудованої моделі кластеризації;
6. скриншот графічного відображення моделі;
7. висновки по роботі.