Лабораторна робота № 2

Знайомство з методами класифікації даних

Мета роботи: Ознайомитися та отримати навички побудови моделей класифікації за допомогою Data Mining GUI бібліотеки WEKA.

Завдання: Побудувати модель класифікації за допомогою методу класифікації.

Загальні відомості

Метод класифікації - це метод аналізу даних, який дозволяє оцінити ймовірність приналежності екземплярів даних до деякого класу залежно від значень їх атрибутів. В якості моделі класифікації рекомендується використовувати структуру даних «дерево» (див. ліст. 1), в якій кожен вузол являє собою точку прийняття рішення на підставі значень атрибутів даних, що класифікуються.

Лістинг 1. Приклад дерева класифікації

На ліст. 1 наведено дерево класифікації, яке дає відповідь на питання «Ви

придбаєте книгу?». У кожному вузлі ставиться конкретизуюче питання (атрибут) з відповідями (значення атрибуту) у гілках, відповідаючи ви переходите до наступного вузла (питання, атрибуту) до тих пір, поки не дійдете до листа зі значенням класу, у прикладі це відповіді «Придбаєте» чи «Не придбаєте» книгу. Перевага класифікаційних дерев полягає у тому, що вони не вимагають надмірної кількості інформації для побудови досить точного та інформативного дерева рішень.

Метод класифікації використовує *відомі* значення атрибутів екземплярів даних та зв'язки між їх значеннями при побудові моделі класифікації. При наявності *нових* екземплярів даних невідомого класу, до даних застосовується раніше побудована модель класифікації і визначається відповідний клас.

Набір даних для WEKA

Набір даних, який буде використаний для прикладу класифікаційного аналізу, містить інформацію, зібрану центром продажу компанії BMW. Центр починає рекламну компанію, пропонуючи розширену дворічну гарантію своїм постійним клієнтам. Подібні компанії вже проводилися, так що центр продажу має у розпорядженні 4500 екземплярів даних щодо попередніх продажів з розширеною гарантією. Цей набір даних охоплює наступні атрибути:

– Розподіл за доходами [0=$0-$30k, 1=$31k-$40k, 2=$41k-$60k, 3=$61k-$75k, 4=$76k-$100k, 5=$101k-$150k, 6=$151k-$500k, 7=$501k+];
– Рік / місяць покупки першого автомобіля BMW;
– Рік / місяць покупки останнього автомобіля BMW;
– Чи скористався клієнт розширеною гарантією?

Файл даних у форматі Attribute-Relation File Format (ARFF) буде виглядати наступним чином, див. ліст. 2.

```
@attribute IncomeBracket
{0,1,2,3,4,5,6,7}
@attribute FirstPurchase numeric
@attribute LastPurchase numeric
@attribute responded {1,0}

@data

4,200210,200601,0
5,200301,200601,1
...
```

Лістинг 2. Файл даних для класифікаційного аналізу у WEKA

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

При побудові моделі класифікації набір даних зазвичай ділять так, щоб частина даних використовувалася для побудови моделі (навчання), а частина - для перевірки її коректності (тестування), щоб переконатися, що модель не є навченою тільки під конкретний набір даних.

Розділіть вибраний набір даних на два файли * .arff (в співвідношенні «2/3» та «1/3» від загальної кількості даних). Завантажте файл «2/3» в програмний пакет WEKA.

Імпортування даних у WEKA

Коли файл з навчальними даними готовий, його потрібно завантажити у WEKA. Запустіть WEKA і виберіть опцію Explorer. У результаті відкриється закладка Preprocess у вікні Explorer. Натисніть на кнопці Open File і виберіть створений вами ARFF-файл. Вікно WEKA Explorer з завантаженими даними показано на рис. 1.

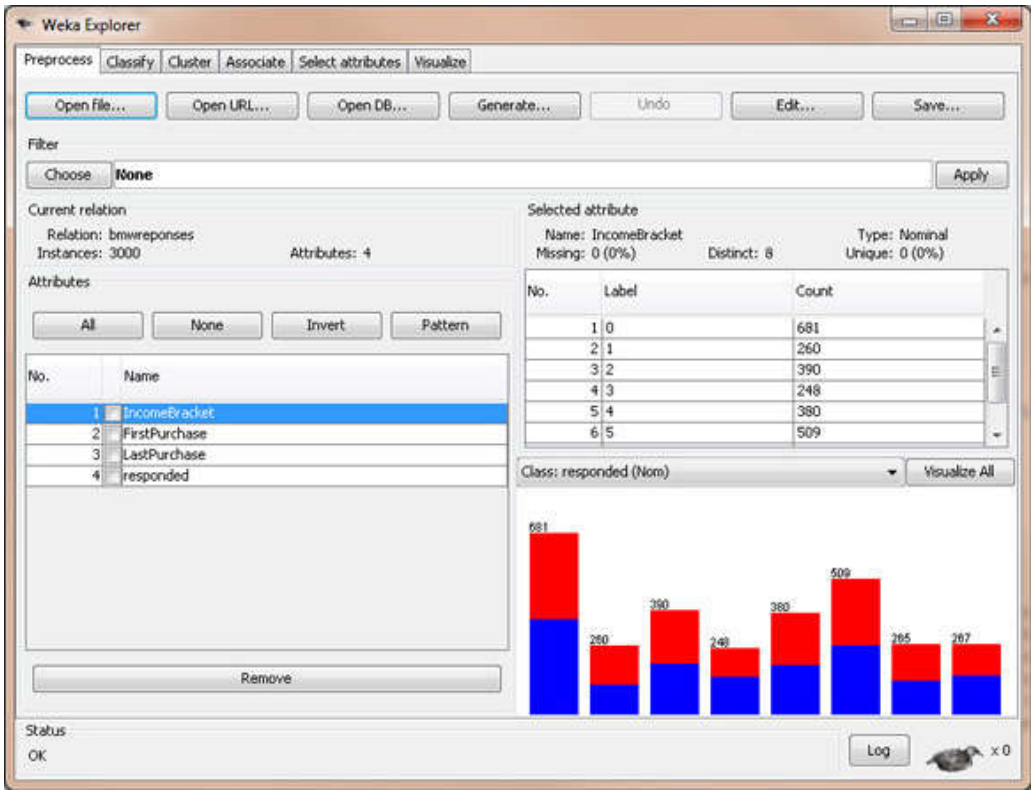

Рис. 1. Дані дилерського центру BMW

У цьому вікні ви можете перевірити дані, на підставі яких ви збираєтесь будувати модель. У лівій частині вікна Explorer показані атрибути даних (Attributes), які відповідають заголовкам стовпців таблиці, а також вказано кількість екземплярів даних (Instances), тобто рядків таблиці. Якщо виділити мишкою однин з заголовків стовпців, тоді в правій частині вікна з’являться значення відповідного атрибуту для різних екземплярів даних. Також існує можливість візуального аналізу даних за допомогою кнопки Visualize All.

Класифікація у WEKA

Виберіть метод класифікації (див. рис. 2): відкрийте закладку Classify, виберіть опцію trees, а потім опцію з відповідним номером Вашої залікової книжки (див. табл. 1).

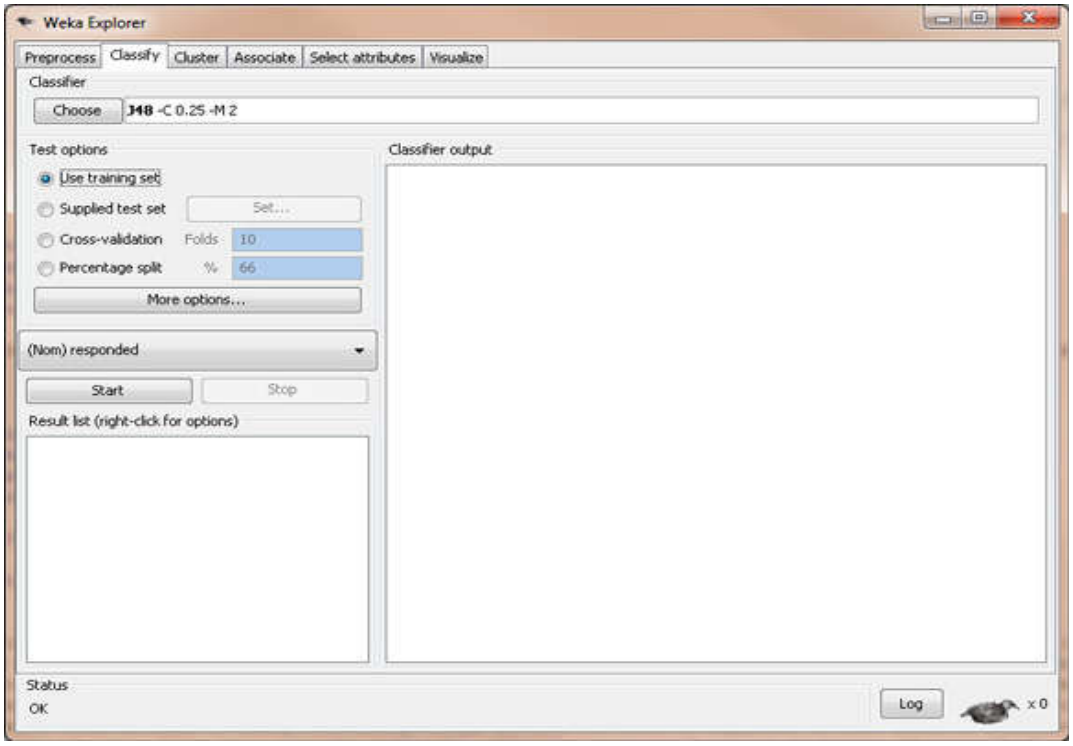

Рис. 2. Вибір методу класифікації даних

Тепер можна розпочати побудову моделі класифікації засобами пакету WEKA. Переконайтеся, що обрана опція Use training set, для того щоб пакет WEKA при побудові моделі використовував саме ті дані, які ви тільки що завантажили з файлу. Натисніть Start. Результуюча модель повинна виглядати так, як показано на ліст. 3.

```
Number of Leaves  :     28
Size of the tree :      43
Time taken to build model: 0.18 seconds
=== Evaluation on training set ===
```

```
=== Summary ===
Correctly Classified Instances        1774               59.1333 %
Incorrectly Classified Instances      1226               40.8667 %
Kappa statistic                          0.1807
Mean absolute error                      0.4773
Root mean squared error                  0.4885
Relative absolute error                 95.4768 %
Root relative squared error             97.7122 %
Total Number of Instances             3000
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.662    0.481      0.587     0.662     0.622      0.616    1
                 0.519    0.338      0.597     0.519     0.555      0.616    0
Weighted Avg.    0.591    0.411      0.592     0.591     0.589      0.616
=== Confusion Matrix ===
   a   b   <-- classified as
1009  516 |   a = 1
 710  765 |   b = 0
```

Лістинг 3. Результат роботи класифікаційної моделі WEKA

На рис. 3 представлена побудована модель класифікації.

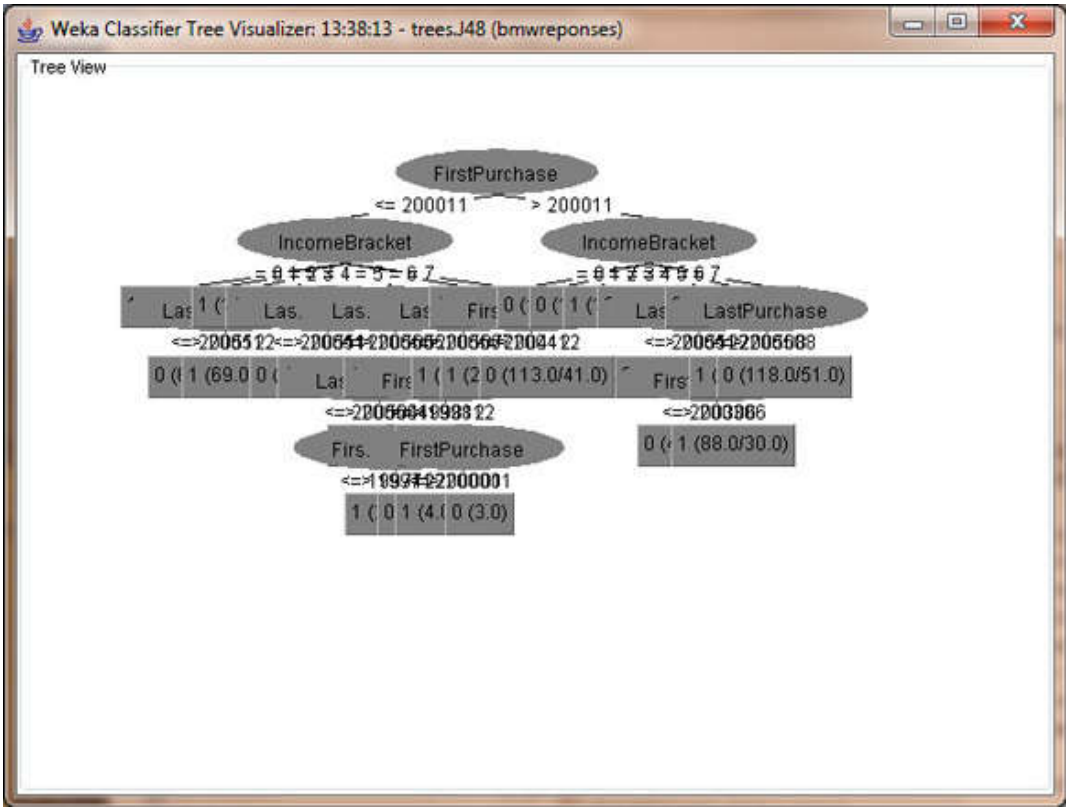

Рис. 3. Графічне відображення моделі класифікації

Модель класифікації (див. рис. 3) побудована на основі навчального набору даних (training set). Цей підхід використовує відомі дані для аналізу зв'язків значень атрибутів навчальних класів та побудови моделі на основі розрахованих зв'язків. Таким чином, коли з'являється новий екземпляр даних (набір значень атрибутів) невідомого класу, потрібно лише провести аналіз значень атрибутів окремого екземпляру даних за допомогою побудованої моделі класифікації та визначити відповідний клас даних.

При побудові моделі класифікації зазвичай використовують набір даних, який ділиться на дві частини. Перша частина (60-80% або 2/3 даних) використовується як навчальний набір для побудови моделі. Після цього тестові дані класифікуються за допомогою побудованої моделі та порівнюються з їх дійсними класами. Такий підхід дозволяє оцінити точність побудованої моделі класифікації.

Тестова перевірка моделі класифікації дозволяє уникнути зайвого

перенавчання моделі. Оскільки модель класифікації будується для класифікації некласифікованих екземплярів, при перевірці її оптимальності використовується тестовий набір даних. Таким чином, гарантується, що побудована модель класифікації зможе з досить високою ймовірністю визначити клас ще некласифікованого екземпляру.

Перевірку моделі треба провести на тестовому наборі даних «1/3» і оцінити, наскільки результати класифікації відрізняються від тестових класів. Для цього в секції Test options виберіть опцію Supplied test set і натисніть Set. Вкажіть файл з даними, що містить тестові «1/3» дані, які не були включені в навчальний набір. При натисканні на кнопку Start, WEKA проведе класифікацію тестових даних і виведе інформацію про оптимальність побудованої моделі (див. рис. 4).

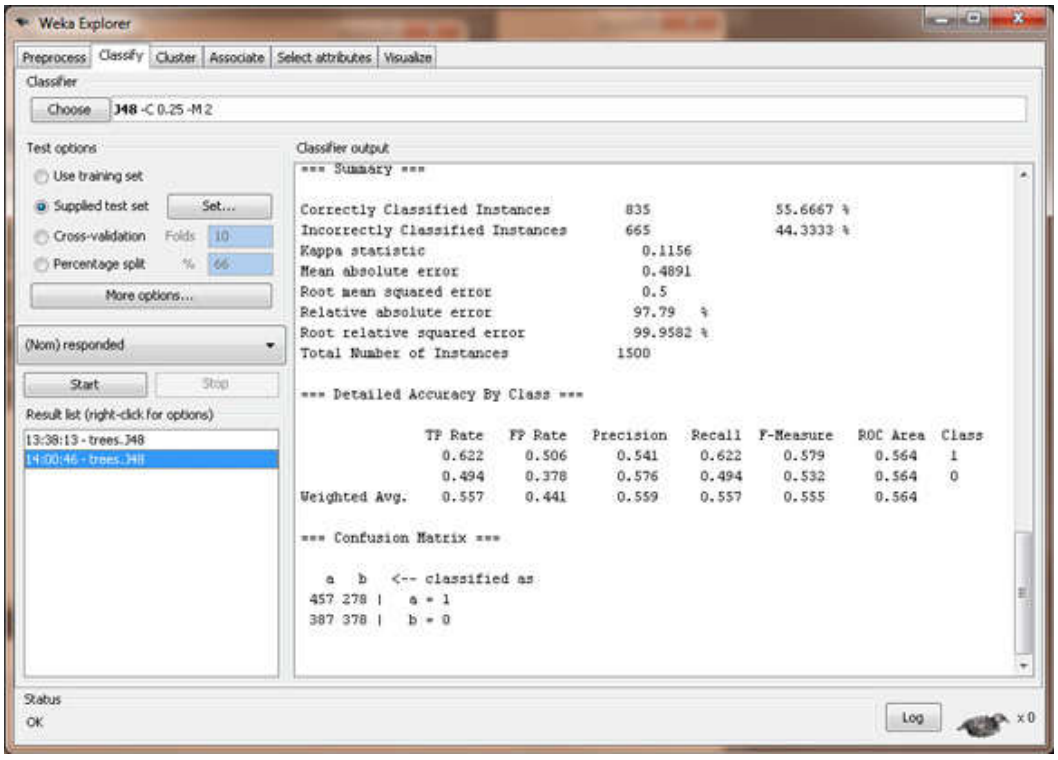

Рис. 4. Перевірка моделі класифікації

Порівнюючи показник Correctly Classified Instances для тестового набору (55,7%) з цим же показником для навчального набору (59,1%), видно, що точність моделі для двох різних наборів даних приблизно однакова. Це означає, що нові

дані, які будуть класифікуватися за допомогою цієї моделі в майбутньому, не зменшать точність її роботи. Для підвищення точності класифікації рекомендується збільшити кількість навчальних і тестових даних, а також число атрибутів класів.

Таблиця 1

Варіанти завдання

Номер	Метод	Номер	Метод
1	Id3	11	NBTree
2	J48	12	REPTree
3	J48graft	13	Id3
4	LADTree	14	J48
5	NBTree	15	J48graft
6	REPTree	16	LADTree
7	J48graft	17	REPTree
8	J48	18	NBTree
9	Id3	19	J48graft
10	LADTree	20	Id3

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 класифікації, скриншот графічного відображення моделі;
6. скриншот перевірки побудованої моделі;
7. висновки по роботі.