МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Інститут Комп'ютерних наук та інформаційних технологій
Кафедра Програмної інженерії та інтелектуальних технологій управління
Спеціальність 122 Комп'ютерні науки
Освітня програма Комп'ютерні науки та інтелектуальні системи

ЛАБОРАТОРНА РАБОТА №1 за курсом

«Інтелектуальний аналіз даних та видобування знань»

Тема лабораторної роботи Попередня обробка даних в WEKA

Виконав студент 5 курсу, групи КН-М422

Захар ПАРАХІН

(підпис, прізвище та ініціали)

Перевірила Оксана ІВАЩЕНКО

(підпис, прізвище та ініціали)

Харків 2022

ЗМІСТ

Вступ

WEKA - бібліотека алгоритмів машинного навчання для вмрішення завдань data mining. Waikato Environment for Knowledge Analysis (WEKA), є вільно поширюваним програмним пакетом з відкритим вихідним кодом для аналізу даних, що реалізований на мові програмування Java.

Система дозволяє використовувати алгоритми до вибіркових даних, надає графічний інтерфейс користувача для роботи з файлами і генерації візуальних результатів.

Weka включає до себе засоби для підготовки обробки даних, класифікації, регресії, кластеризації, вибору властивостей, пошуку асоціативних правил і візуалізації. Система також має можливість для розробки нових підходів до інтелектуального аналізу даних, або використання нетипових як, наприклад, Memory-based reasoning.

1 Хід виконання роботи

1.1 Підготовка середовища до роботи

Для виконання лабораторної работи було встановлено середовище WEKA. При запуску середовища відкривається вікно вибору графічного інтерфейсу користувача для різних задач (рис. 1).

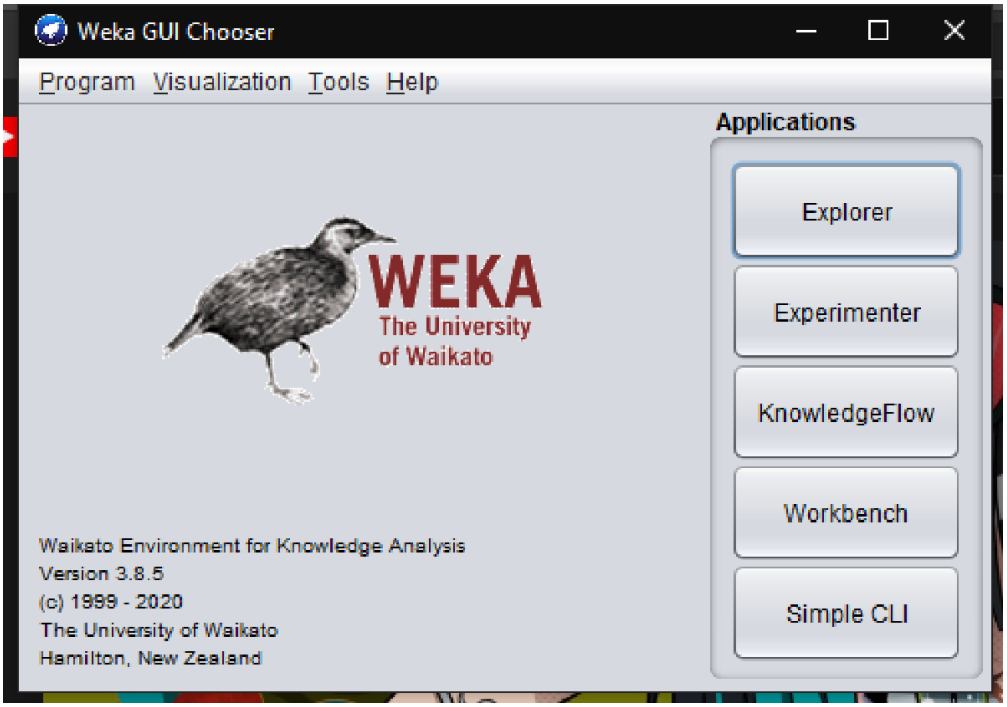

Рисунок 1 - Вікно вибору інтерфейсу

У початковому вікні надається доступ до п’яти модулів програми серед яких (окрім Workbench):

Explorer - середовище для дослідження даних;

Experiment - середовище для проведення порівняльного аналізу різних алгоритмів при обробці одного набору даних;

KnowledgeFlow - середовище аналогічне до Explorer, але з імплементацією інтерфейса drug-and-drop і підтримкою інкрементного навчання;

SimpleCLI - командний інтерфейс для прямого виконання команд.

Головне же меню складається з чотирьох пунктів:

1. **Program**

   LogWindow - вікно логів, зберігає всю інформацію в потоках вводу-виводу;

   Memory usage - використання пам’яті;

   Settings - налаштування програми;

   Exit - вихід;

2. **Visualization** - група засобів візуалізації

   Plot - відображення двовимірного графіку набору даних;

   ROC - відображення раніше збереженної ROC-кривої;

   TreeVisualizer - відображення направлених графів (дерев рішень);

   GraphVisualizer - візуалізація графіку у форматах XML BIF або DOT для Байесових мереж (Bayers nets);

   BoundaryVisualizer - візуалізовує границі рішень класифікаторів у двох вимірах;

3. **Tools**

   ArffViewer - редактор artff-файлів;

   SqlViewer - модуль перегляду баз даних, для SQL запитів;

   Bayers net editor - модуль для редактування, візуалізації та навчання Байесових мереж;

4. **Help** - онлайн ресурси для вивчення WEKA;

1.2 Ознайомлення з базовими операціями попередньої обробки даних

1.2.1 Завантаження даних (Preprocess)

Ця частина програми надає доступ для завантаження і початкової обробки даних. Усього є декілька варіантів завантаження даних:

- з файлу;

- згенерувати дані моделі;
- з URL адреси;
- з бази даних

В останіх двох необхідно просто вказати посилання, тому більше буде розглянутий первший варіант. При відкритті файлу, є можливість у списку вибрати розширення файлів (arff, CSV, C4.5, libsvm, бінарні). Arff - це текстовий файл, який описує список об’єктів з спільними атрибутами, і складається структурно з заголовку (метадані і ім’я атрибутів) і даних (значень об’єктів після тега @data).

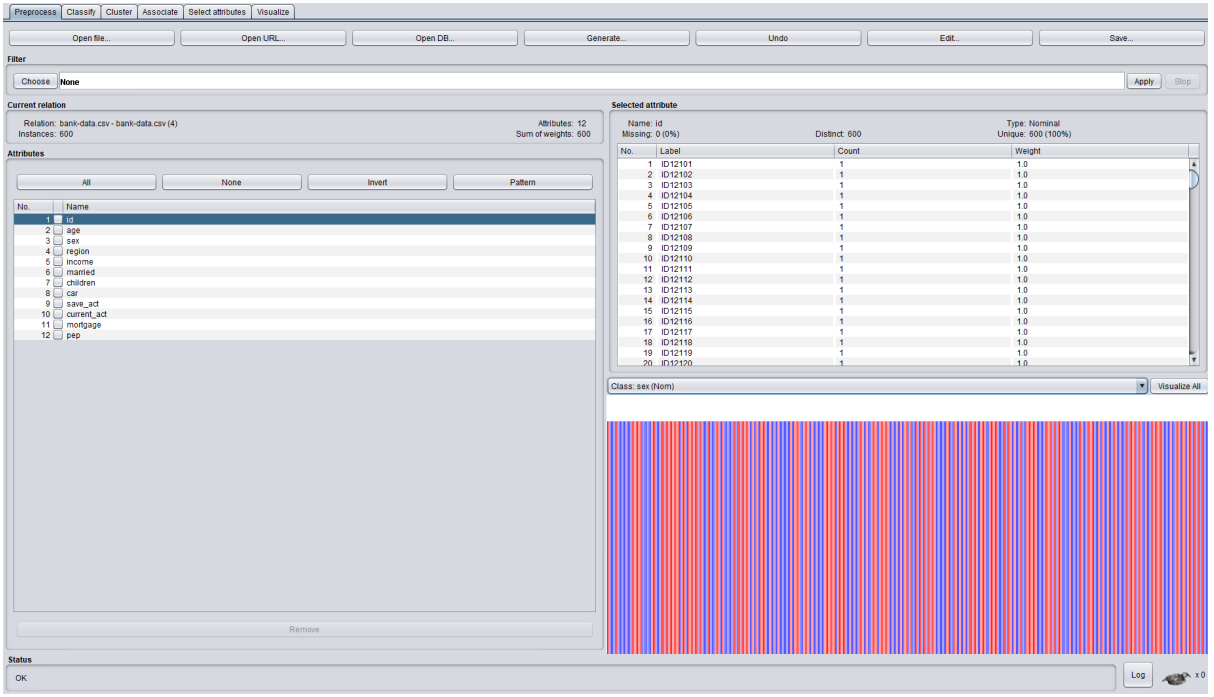

Рисунок 2 - Відображення завантаженої множини даних

“Edit” - для редактування вхідних даних, що відкриває вікно “Viewer”.

1.2.2 Вибір та фільтрація атрибутів

Перед видобуванням даних потрібно позбавитися від непотрібних атрибутів у наших записах. Наприклад, використання id є для нас беззмістовним так як при селекції з декількома атрибутами не змінює і є фактично фіктичною змінною при аналізі, яка в непідготовленому варіанті мала сенс як індифікатор.

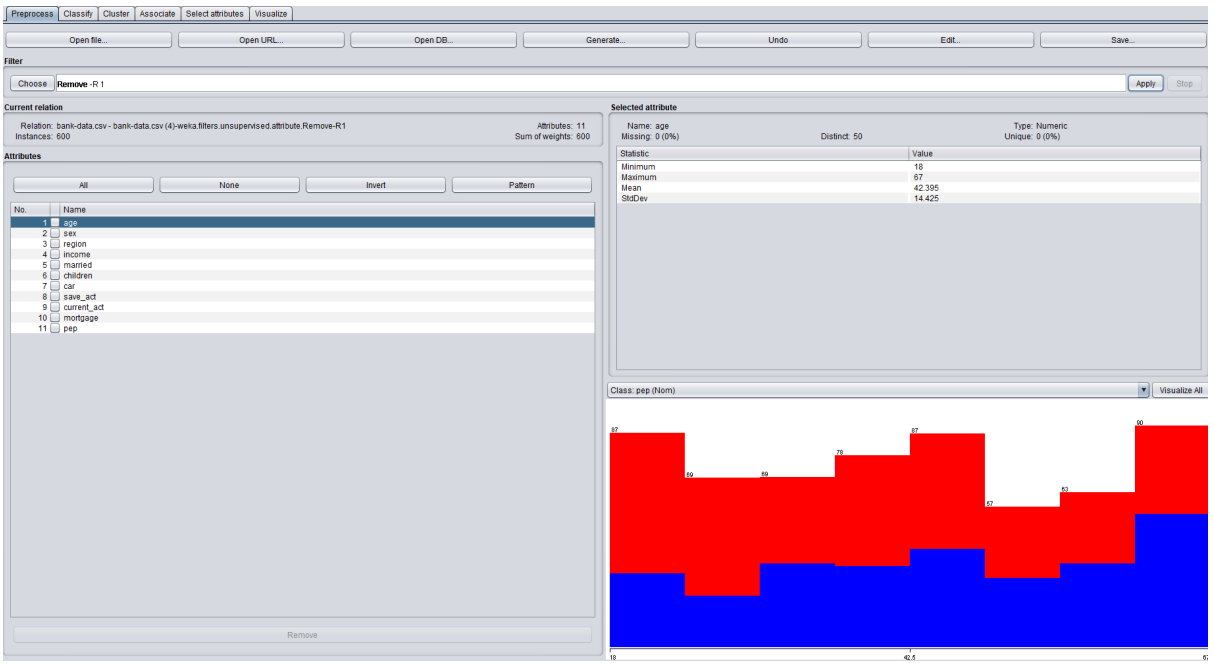

Рисунок-3 Зміни у множині даних після видалення атрибуту id

В програмі можна використовувати селектори за атрибутами, які комбінуються, також за ними можна видаляти певні частини як id, а отриману множину даних зберігати у зручному форматі arff.

Вказаний файл має наступний вміст:

```
@relation bank-data.csv-weka.filters.unsupervised.attribute.Remove-R1

@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data

48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES
40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES
```

Рисунок 4 - Вміст файлу arff

У заголовку записуються ім’я даних і їхні метадані (як типи). Всі імена мають бути унікальними і відповідати порядку кожного з атрибутів з вказаним ім’ям і типом обов’язково. Також типом даних можуть бути категоріальні перечні (можливі значення як {Yes, NO})Наприклад:

% коментар

@relation bank-data.cvs-weka.filters.unsupervised.attribute.Remove-R1

@attribute age numeric

В другій частині файлу знаходяться самі дані, перелічені через кому. Для кожного запису використовується окрема строка. Та дані можуть мати пропущені значення, які позначаються знаком питання (?). Строчні дані разом з розділовими символами беруться в кавички.

1.2.3 Discretization

Фільтр екземплярів, який дискретизує діапазон числових атрибутів у наборі даних на номінальні атрибути. Дискретизація здійснюється за допомогою простого групування. Пропускає атрибут класу, якщо встановлено.

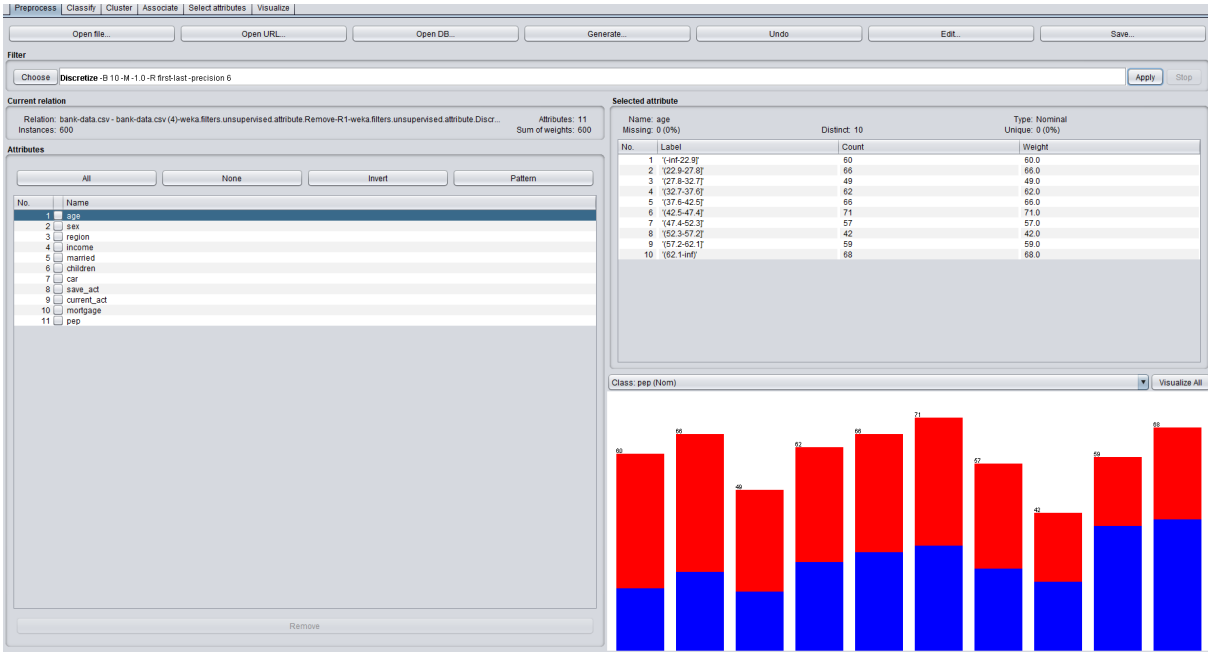

Рисунок 5 - Використання фільтру Discretization

1.2.4 Пропущені дані

Перевіримо дані на пропущені дані. Вибравши атрибут, ми бачимо інформацію про нього – значення якого типу він містить, скільки пропущених значень тощо. Перевіривши всі атрибути, ми бачимо, що пропущених значень.

Для того, щоб заповнити пропуски в WEKA існує 2 фільтри: ReplaceMissingValues, який заповнює пропуски модою та середнім значенням та ReplaceMissingWithUserConstant, який заповнює пропуски заданим користувачем значенням. Скористайтеся фільтром ReplaceMissingValues. Знайти цей фільтр можна натиснувши Choose… і вибравши weka→filters→unsupervised→attribute→ReplaceMissingValues. Застосуємо цей фільтр, натиснувши Apply.

1.3 Аналіз датасету

В WEKA є можливості аналізу даних, візьмемо хоча б найпростіші статистичні методи як регресійні лінійні моделі, як приклад, для цього потрібно перейти в Classify і обрати методи, регресійні моделі знаходяться в function, візьмемо SimpleLogic, для income атрибуту буде сформована модель.

```
Class '(22448.977-28260.566]' :
1.25 +
[age='(-inf-22.9]'] * -8.98 +
[age='(22.9-27.8]'] * -0.88 +
[age='(27.8-32.7]'] * 0.96 +
[age='(32.7-37.6]'] * 0.42 +
[age='(37.6-42.5]'] * 0.22 +
[age='(47.4-52.3]'] * -0.55 +
[age='(52.3-57.2]'] * 0.08 +
[age='(57.2-62.1]'] * -0.04 +
[age='(62.1-inf)'] * -0.58 +
[region=INNER_CITY] * -0.34 +
[region=TOWN] * -0.02 +
[region=RURAL] * 0.23 +
[region=SUBURBAN] * 0.05 +
[married=YES] * -0.09 +
[children='(-inf-0.3]'] * -0.14 +
[children='(0.9-1.2]'] * 0.25 +
[children='(1.8-2.1]'] * 0.11 +
[children='(2.7-inf)'] * -0.62 +
[car=YES] * 0.14 +
[save_act=YES] * -0.51 +
[current_act=YES] * -0.33 +
[mortgage=YES] * 0.27 +
[pep=NO] * 0.48
```

Рисунок 7 - ЛРМ для income

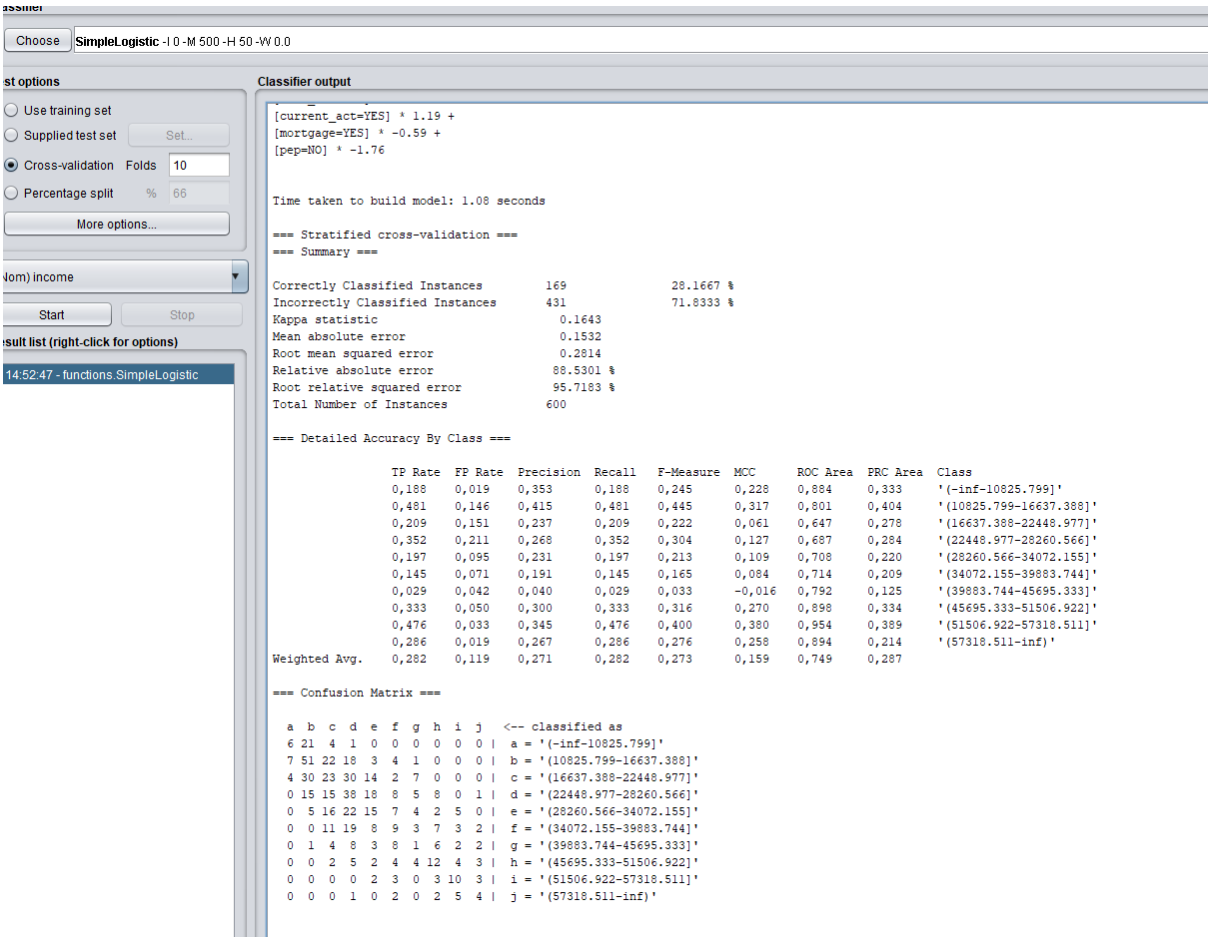

Рисунок 8 - Результати моделі

Однак, можливості інтелектуального аналізу даних не обмежуються визначенням одного параметра. Основне завдання аналізу - виявлення залежностей і зв'язків у великих наборах даних. Інтелектуальний аналіз, як правило, використовується не для того, щоб визначити яке-небудь конкретне значення, а для того, щоб побудувати модель, що дозволяє аналізувати зв'язки між даними, прогнозувати результати і робити обґрунтовані висновки, які підтверджуються зібраними статистичними даними.

Висновки

Було вивчено базові основні функції і засоби, які надає WEKA, для аналізу даних. Ця програма надає прості методи для підготовки даних, і подальшого їх аналізу відповідно до типових потреб, як кластеризація, побудова регресійних моделей, побудова дерев рішень.

Для використання даних дані потрібно або додати через файл з одним з розширень, або для віддаленого через URL. Окрім цього, пропонується збереження результатів роботи в форматі .arff, який розділяє метадані і самі дані сутностей аналізу, що додаткова надає можливість до маніпуляції даних додаванням нових сутностей, разом з цим є можливість знаходити пропущені дані і при допомозі фільтрів виправляти.

Окремо треба зауважити спрощеність використання методів, як і для фільтрації певних даних за атрибутами чи властивостями так і для вибору певного методу як регресій або кластеризації, чи побудови дерев, що значно спрощує задачі аналізу, а результати одразу виводяться.

Також, важливою складовою програми є засоби візуалізації, що одразу надають просте відображення підготовлених даних і вже потім при використанні певних методів інтелектуального аналізу.

Всі необхідні маніпуляції з даними були виконані для опрацювання навчального матеріалу. Проблеми виникали тільки на етапі підключення файлу, через ігнорування програмою роздільних символів для формату CSV, але були вирішені через інструменти програми.

Список джерел інформації

1 Data Mining: Practical Machine Learning Tools and Technique \ Ian Witten, Eibe Frank, Mark Hall

2 WEKA. Руководство по использованию \ Хабр online ресурс https://habr.com/ru/post/590565/

3 Інтелектуальний аналіз даних: Навчальний посібник \ А. О. Олійник, С. О. Субботін, О. О. Олійник