

THỰC HÀNH KHAI PHÁ DỮ LIỆU

Bài 1. Thực hành trên weka

Giáo viên: TS. Trần Mạnh Tuấn

Bộ môn: Hệ thống thông tin

Khoa: Công nghệ thông tin

Email: tmtuan@tlu.edu.vn

Điện thoại: 0983.668.841

Nội dung

1

Giới thiệu về Python

2

Giới thiệu về R

3

Giới thiệu về weka

4

Tiền xử lý dữ liệu trên weka

➤ GIỚI THIỆU VỀ PYTHON

- Python có thể là cửa ngõ để mọi người bước vào thế giới lập trình máy tính, và là một phương tiện để bạn nhận được khoản tiền lương béo bở đi kèm với một công việc đầy sáng tạo và nhiều niềm vui.
- Được đặt theo tên một nhóm hài kịch và nổi tiếng với cú pháp đơn giản và thanh lịch, Python được sử dụng cho nhiều loại ứng dụng từ các trò game đơn giản đến các thuật toán tìm kiếm phức tạp
- Python luôn nằm trong top 10 ngôn ngữ lập trình phổ biến nhất ở tất cả các bảng xếp hạng lớn
- Python có tốc độ rất nhanh, mạnh mẽ và có mặt ở khắp mọi nơi.

➤ GIỚI THIỆU VỀ PYTHON

- Python là 1 ngôn ngữ thông dịch, tương tác động.
- Python là ngôn ngữ dễ học: cú pháp đơn giản, rõ ràng sử dụng một số lượng không nhiều các từ khoá
- Python là ngôn ngữ dễ hiểu: Mã lệnh dễ đọc và dễ hiểu.
- Python có tương thích cao (highly portable): chạy trên nhiều nền tảng hệ điều hành khác nhau bao gồm Windows, Mac OSX và Linux.
- Python hỗ trợ hướng đối tượng – Python hỗ trợ mô hình lập trình hướng đối tượng và các kỹ thuật lập trình đóng gói mã nguồn bên trong các đối tượng.
- Python hỗ trợ Module hóa

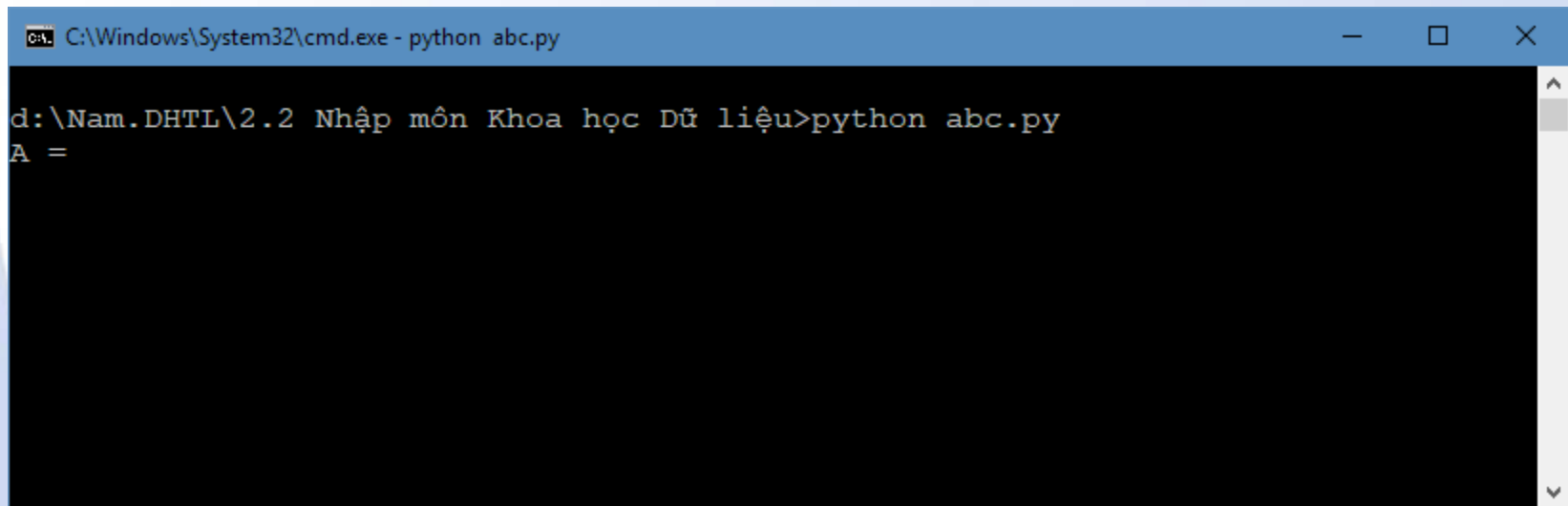
➤ GIỚI THIỆU VỀ PYTHON

Cài đặt



➤ GIỚI THIỆU VỀ PYTHON

- Python có 2 chế độ thực thi:
 - Chế độ thực thi: chỉ ra chương trình cần thực hiện
 - Chế độ dòng lệnh
- Chế độ thực thi: “python abc.py” chạy file “abc.py”



```
C:\Windows\System32\cmd.exe - python abc.py

d:\Nam.DHTL\2.2 Nhập môn Khoa học Dữ liệu>python abc.py
A =
```

➤ GIỚI THIỆU VỀ PYTHON

- Khai báo biến trong Python
- Từ khóa
- Định danh
- Các kiểu dữ liệu cơ bản trong Python
- Các thao tác với file dữ liệu
- Lỗi và truy vết lỗi
- Tham khảo:
[“http://vietjack.com/python/index.jsp”](http://vietjack.com/python/index.jsp)

➤ GIỚI THIỆU VỀ PYTHON

Làm việc với file trong python:

- Mở file: `doi_tuong_file = open(ten_file [, access_mode][, buffer])`
- Đóng file: `fileObject.close()`
- Đọc file: `doi_tuong_file.read(giatri)`
- Ghi file: `doi_tuong_file.write(string)`
- Xóa file: `os.remove(ten_file)`

➤ GIỚI THIỆU VỀ PYTHON

Để thực hành khai phá dữ liệu hay học máy trên python:

- Cài đặt thư viện Sklearn
- Cài đặt chung trên ứng dụng Anaconda.

<https://docs.anaconda.com/anaconda/install/windows>

➤ GIỚI THIỆU VỀ R

- ❖ Cài đặt R:

<http://cran.r-project.org/>

- ❖ Cài đặt các gói cần thiết

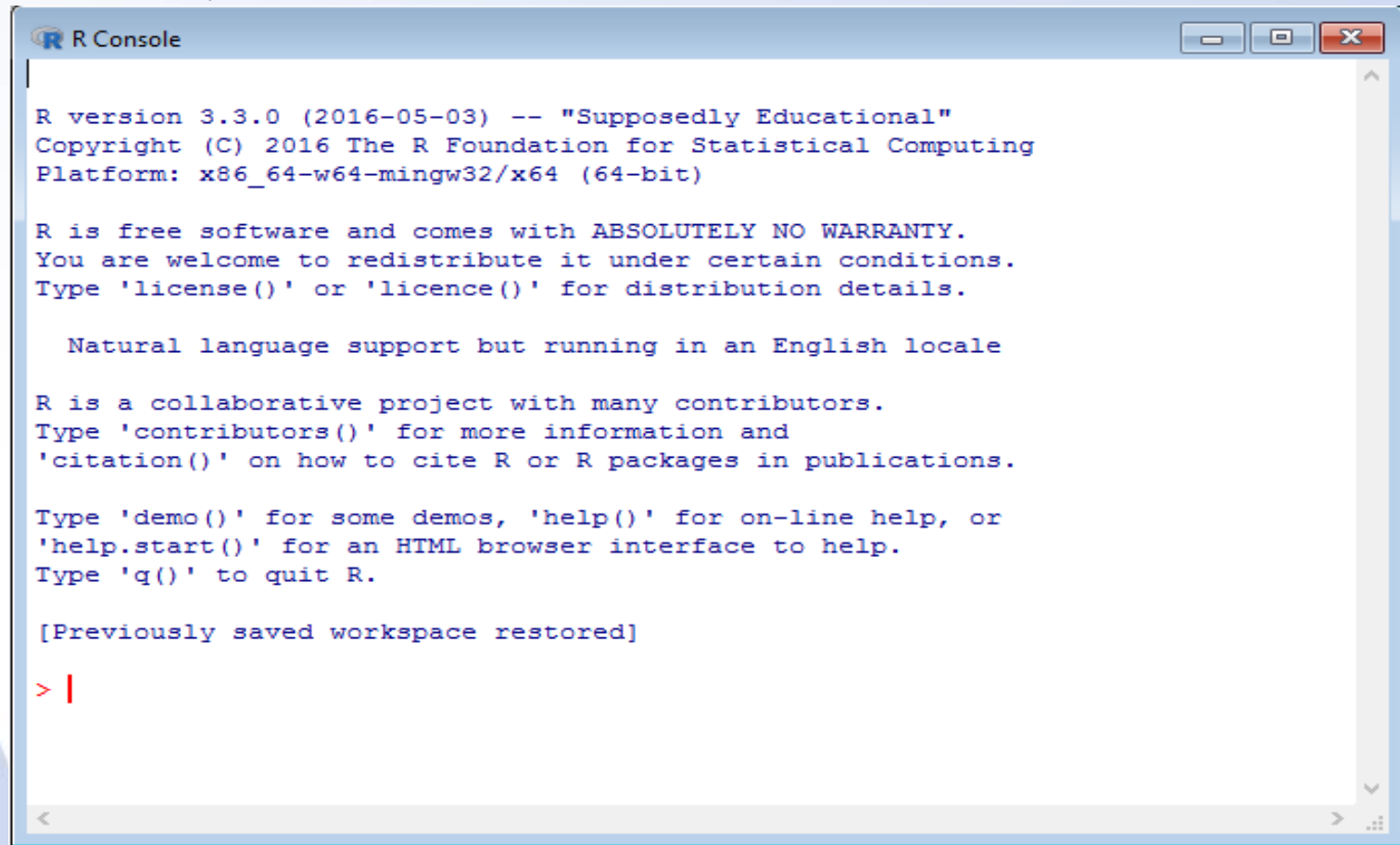
`install.packages("tên gói")`

Hoặc chọn **Packages\Install packages.**

- ❖ Thư mục làm việc (work directory) của R:
- ❖ Kiểm tra bằng lệnh `getwd()`

➤ GIỚI THIỆU VỀ R

□ Giao diện



```
R Console

R version 3.3.0 (2016-05-03) -- "Supposedly Educational"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

    Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

➤ GIỚI THIỆU VỀ R

- **Nhập dữ liệu trực tiếp**

```
> luong=c(12,14,13,15,12,14,13,15,13,14,15,18)
```

- **Điều chỉnh dữ liệu: edit**
- **Đọc dữ liệu từ file: read.table, read.csv, ...**

➤ GIỚI THIỆU VỀ R

- **Nhập dữ liệu trực tiếp**

```
> luong=c(12,14,13,15,12,14,13,15,13,14,15,18)
```

- **Điều chỉnh dữ liệu: edit**

- **Đọc dữ liệu từ file: read.table, read.csv, ...**

```
Auto=read.table("http://www-bcf.usc.edu/~gareth/ISL/Auto.data")
```

➤ GIỚI THIỆU VỀ R

Các gói hỗ trợ đọc dữ liệu trên R

Tên gói	Cài đặt packages	Công dụng
xlsx	<code>install.packages("xlsx")</code>	Đọc data đuôi .xlsx (file Excel)
gdata	<code>install.packages("gdata")</code>	Đọc data đuôi .xls (file Excel)
foreign	<code>install.packages("foreign")</code>	Đọc data đuôi .sav (file SPSS), .dta (file Stata) ...
hexView	<code>install.packages("hexView")</code>	Đọc data đuôi .wf1 và .WF1 (file Eviews)

➤ GIỚI THIỆU VỀ R

Đọc dữ liệu từ file

- ✓ Kiểm tra số chiều (dim), tên biến (names)
- ✓ Loại bỏ các dòng có dấu ? (missing data)
`Auto=na.omit(Auto)`
- ✓ Truy xuất đến các biến: `tên_file$tên_biến`
(hoặc sử dụng `attach(tên_file)` – không cần \$)
- ✓ Khi đó các biến được coi như các đối tượng khác

➤ GIỚI THIỆU VỀ WEKA

✓ Phần mềm weka:

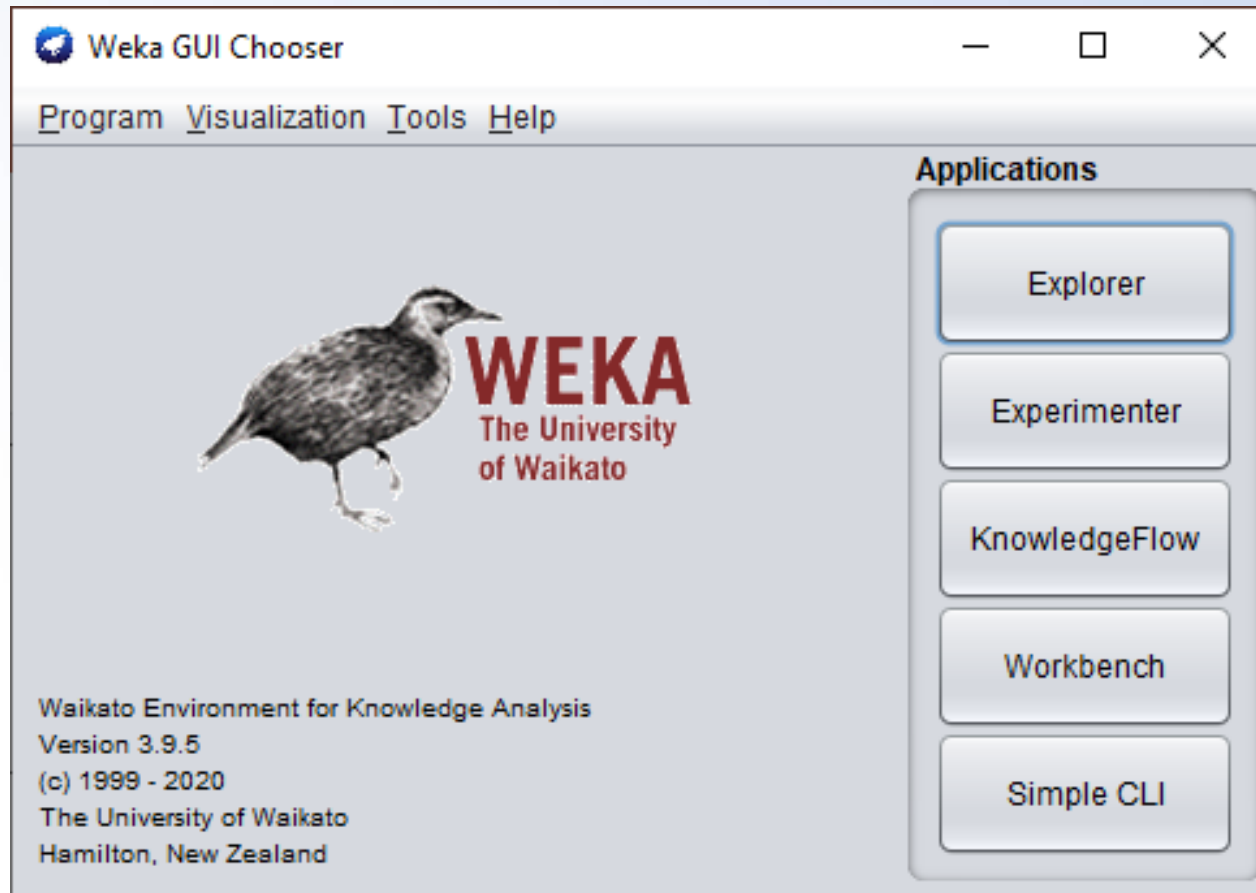
<https://www.cs.waikato.ac.nz/ml/weka/downloading.html>

✓ Kèm theo java sdk:

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

➤ GIỚI THIỆU VỀ WEKA

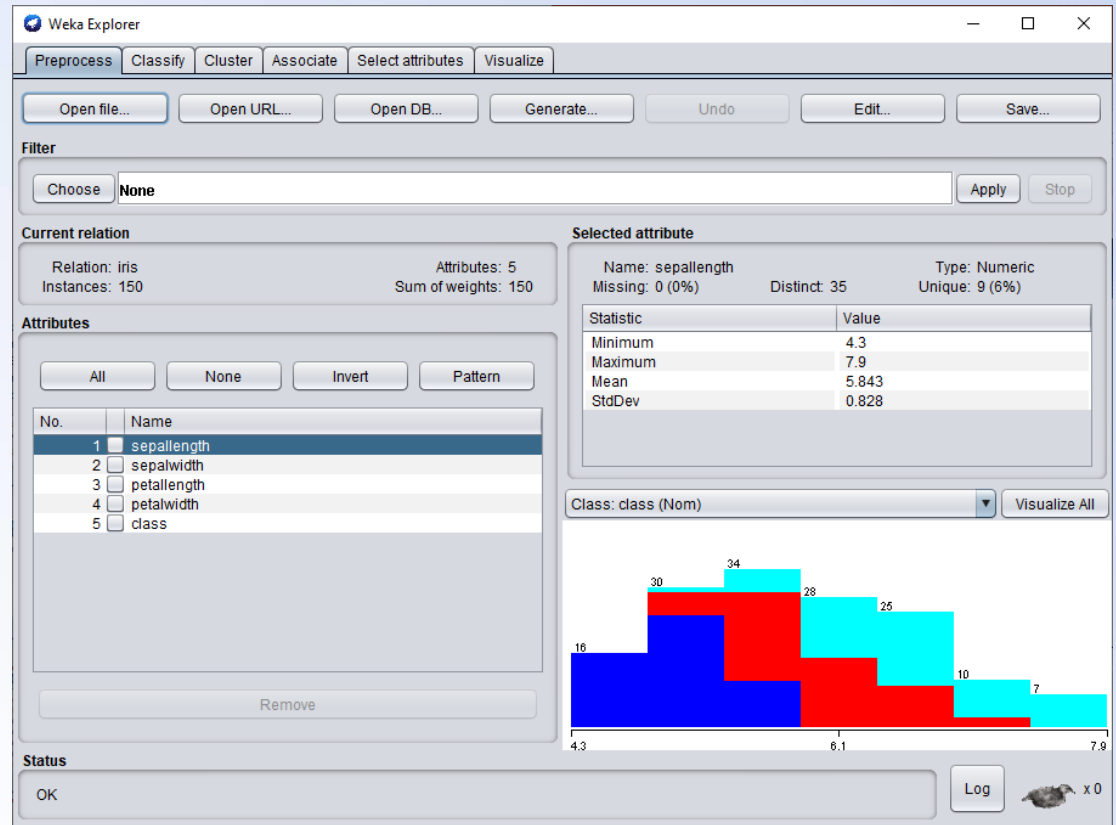
✓ Giao diện weka 3.9.5:



➤ GIỚI THIỆU VỀ WEKA

Chức năng explorer chính của weka, thao tác cơ bản:

- ✓ Tiền xử lý dữ liệu
- ✓ Phân lớp
- ✓ Phân cụm
- ✓ Khai phá luật kết hợp
- ✓ Lựa chọn thuộc tính
- ✓ Trực quan hóa



➤ GIỚI THIỆU VỀ WEKA

Chức năng Experimenter:

- ✓ Thiết kế các thí nghiệm
 - ✓ Lựa chọn thuật toán và tập dữ liệu
 - ✓ Chạy thí nghiệm
 - ✓ Phân tích kết quả (so sánh các kết quả,...)
-

➤ GIỚI THIỆU VỀ WEKA

Chức năng KnowlegeFlow:

- ✓ Thiết kế quá trình khai phá dữ liệu 1 cách trực quan
- ✓ Từ xử lý dữ liệu -> chạy mô hình -> trình bày kết quả

➤ GIỚI THIỆU VỀ WEKA

Chức năng Workbench:

- ✓ Tổng hợp các chức năng ở trên vào trong một ứng dụng
- ✓ Cung cấp cho người sử dụng công cụ mạnh để khai phá dữ liệu

➤ **GIỚI THIỆU VỀ WEKA**

Chức năng Simple CLI:

- ✓ Cho phép người dùng tương tác với WEKA bằng cách gõ lệnh

➤ Tiền xử lý dữ liệu trên weka

Tập tin ARFF:

- ✓ Mô tả đối tượng trong không gian n- chiều
 - ✓ Tập tin ARFF có phần header
 - ✓ Tập tin ARFF có phần data
 - ✓ Các kiểu dữ liệu
-

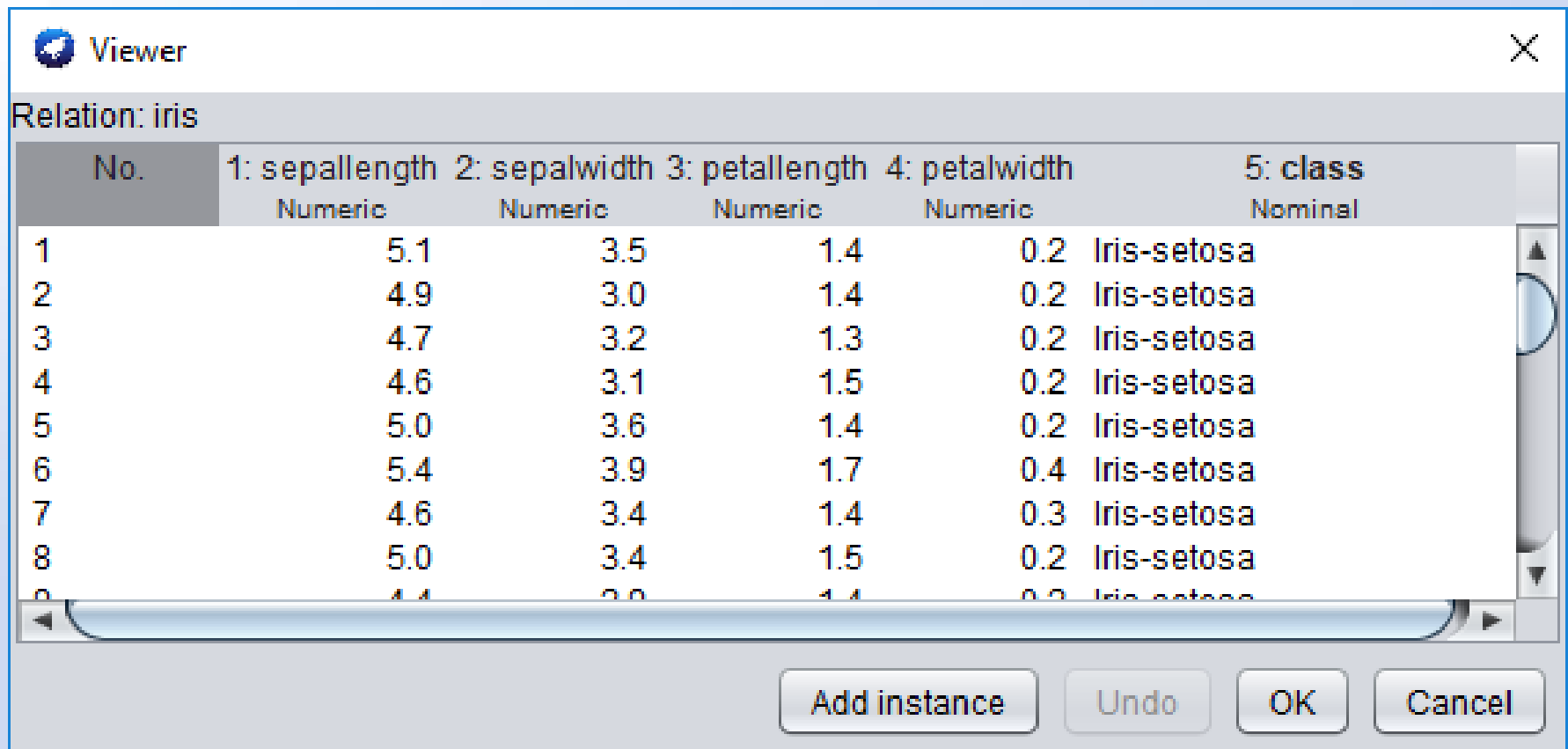
➤ Tiền xử lý dữ liệu trên weka

Tập tin ARFF:

- ✓ Là một văn bản theo bảng mã ASCII
 - ✓ Mô tả các đối tượng có cùng chung tập thuộc tính
 - ✓ Được sử dụng làm định dạng chuẩn cho dữ liệu được dùng bởi các mô hình của weka
-

Tiền xử lý dữ liệu trên weka

Tập tin ARFF:



Relation: iris

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa

Add instance Undo OK Cancel

➤ Tiền xử lý dữ liệu trên weka

Tiền xử lý dữ liệu trong weka preprocessing

- Đọc dữ liệu dưới nhiều hình thức
- Hiệu chỉnh dữ liệu
- Biểu diễn thông tin về tập dữ liệu
- Lưu trữ dữ liệu

➤ Tiền xử lý dữ liệu trên weka

Đọc dữ liệu preprocessing

- Đọc dữ liệu từ file: open file
- Đọc dữ liệu từ địa chỉ URL: open URL
- Đọc dữ liệu từ 1 CSDL: open DB
- Đọc dữ liệu phát sinh (phát sinh dữ liệu từ các bộ phát sinh dữ liệu DataGenerators): Generators

➤ **Tiền xử lý dữ liệu trên weka**

Hiệu chỉnh dữ liệu

- Edit: biểu diễn dữ liệu dưới dạng bảng.
- Nhấn chuột phải ra các chức năng weka hỗ trợ tiền xử lý dữ liệu.

➤ Tiền xử lý dữ liệu trên weka

Mô tả chức năng Preprocessing

➤ Các thao tác trên dữ liệu:

Get mean	<u>Lấy trung bình giá trị của 1 thuộc tính trong tất cả các mẫu</u>
Set all values to	<u>Đặt giá trị của 1 thuộc tính trong tất cả các mẫu bằng giá trị do người dung định</u>
Set missing values to	<u>Đặt giá trị thiếu của 1 thuộc tính bằng giá trị do người dung định</u>
Replace values with	<u>Thay thế giá trị cũ của 1 thuộc tính bằng giá trị mới</u>

➤ Tiền xử lý dữ liệu trên weka

Mô tả chức năng Preprocessing

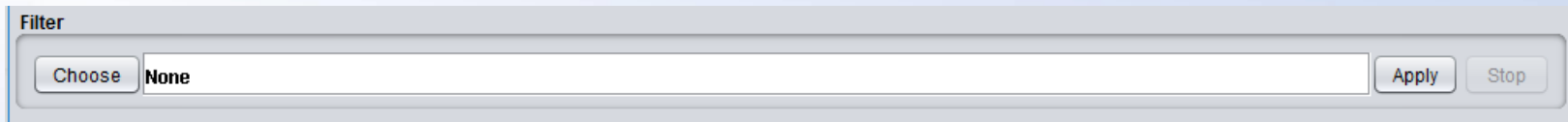
➤ Các thao tác trên dữ liệu:

Rename attribute	<u>Đổi tên thuộc tính</u>
Attribute as class	<u>Đặt thuộc tính đang chọn làm thuộc tính phân lớp</u>
Delete attribute	<u>Xóa thuộc tính</u>
Sort data	<u>Sắp xếp dữ liệu theo thuộc tính được chọn</u>
Optimal column width	<u>Tối ưu hóa chiều rộng cột</u>

➤ Tiền xử lý dữ liệu trên weka

Mô tả chức năng Preprocessing

➤ Lọc dữ liệu:



➤ Tiền xử lý dữ liệu trên weka

Mô tả chức năng Preprocessing

- Biểu diễn thông tin về tập dữ liệu:
 - Cung cấp thông tin về tập dữ liệu (Current relation): tên quan hệ, số mẫu, số thuộc tính
 - Danh sách các thuộc tính (Attribute): danh sách các thuộc tính có thứ tự
 - Thông tin về 1 thuộc tính chọn (selected attribute): tên thuộc tính, tỉ lệ thiếu, loại dữ liệu, các giá trị và số lần xuất hiện
 - Biểu diễn trực quan từng thuộc tính (class)

➤ **Tiền xử lý dữ liệu trên weka**

Tiền xử lý dữ liệu bán tự động

- Người dùng có thể sử dụng các chức năng của Preprocessing để điều chỉnh từng giá trị của thuộc tính
- Convert từ file *.csv sang file *.arff
- Convert từ file *.txt sang file *.arff
- Thực hành trên bộ dữ liệu: tính các độ đo cho các trường

Trao đổi, câu hỏi?