Báo cáo Lab 01 Preprocessing Data Mining

Giáo viên hướng dẫn :

* Lê Hoài Bắc
* Nguyễn Ngọc Đức
* Nguyễn Duy Khánh
* Nguyễn Khánh Toàn

1. **Giới thiệu chung:**

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ Tên** | **Công việc** | **% đóng góp**  **(tối đa 100%)** |
| 18120647 | Lê Thanh Viễn | Cài đặt weka, Đọc dữ liệu vào Weka | 50% |
| 18120363 | Đặng Văn Hiển | Khám phá tập dữ liệu Weather, Khám phá tập dữ liệu Tín dụng Đức, viết báo cáo | 50% |

1. A screenshot of a computer

   Description automatically generated with medium confidence**Cài đặt Weka:**
2. Ý nghĩa các nhóm điều khiển:

* Current relation : Thông tin về tập dữ liệu hiện tại như tên, số lượng thuộc tính, số lượng mẫu bản ghi.
* Attributes : Danh sách các thuộc tính.
* Selected attribute : Thông tin thuộc tính được chọn như tên, loại, số lượng và tỉ lệ thiếu, số lượng và tỉ lệ giá trị duy nhất. Với bảng đại lượng thống kê với Numeric và liệt kê nhãn với Nominal.

1. Ý nghĩa 5 tab trong giao diện Explorer:

* Preprocess : Đưa dữ liệu vào và tiền xử lý.
* Classify : train model bài toán phân loại và test model (Bayes, hồi quy, cây quyết định, mạng perceptron nhiều lớp,.. )
* Cluster : dùng thuật toán phân nhóm (phân nhóm theo tầng, K-means,..)
* Associate : khám phá các luật kết hợp.
* Select Attributes : chọn các thuộc tính quan trọng.
* Visualize : ma trận biểu đồ trực quan giữa mỗi 2 thuộc tính.

1. **Làm quen với Weka:**
   1. **Đọc dữ liệu vào Weka : (tập dữ liệu breast\_cancer.arff)**

**Shape, rectangle

Description automatically generated with medium confidence**

1.Tập dữ liệu có 286 mẫu.

2.Tập dữ liệu có 10 thuộc tính.

3.Thuộc tính được dùng làm lớp là Class. Thay đổi thuộc tính dùng làm lớp là :

Graphical user interface

Description automatically generated

4.Có 2 thuộc tính bị thiếu dữ liệu là node-caps và breast-quad. Với node-caps thiếu dữ liệu nhiều nhất 8 giá trị (3%) và thiếu ít nhất là breast-quad 1 giá trị (3%)

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

-Những cách giải quyết vấn đề missing values:

+ Bỏ qua những bản ghi với dữ liệu thiếu

+ Điền dữ liệu thiếu thủ công

+ Sử dụng hằng toàn cục như (NULL, unknown,...)

+ Điền với giá trị thuộc tính trung bình hoặc được suy diễn phù hợp (phương pháp cây quyết định hay bộ phân lớp Bayesian)

5.Tên biểu đồ cột số người ung thư vú theo kích thước khối u và có không tái phát u

- Màu xanh là không tái phát và màu đỏ là có tái phát u .

- Đồ thị biểu diễn số người ung thư vú theo kích thước khối u với đoạn màu đỏ là những người đã tái phát và xanh là chưa xảy ra tái phát.

Chart, bar chart

Description automatically generated

* 1. **Khám phá tập dữ liệu Weather :**

**Rectangle

Description automatically generated with low confidence**

1.Tập dữ liệu có 14 mẫu, 5 thuộc tính. Thuộc tính lớp là play

Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric) :

+ Categorical: outlook, windy, play

+ Numeric: temperature, humidity

2.Weka không cung cấp đầy đủ 5 giá trị trên ( chỉ cung cấp Min và Max )

Table

Description automatically generated

- temperature:

min=64

max=85

Q1, median, Q3 = ?

- humidity:

min=65

max=96

Q1, median, Q3 = ?

Graphical user interface, application, PowerPoint

Description automatically generated3 . Xem các thuộc tính khác của dataset dưới dạng đồ thị

A picture containing graphical user interface

Description automatically generated4 . Thuật ngữ là scatter plot. Cặp thuộc tính tương quan humidity- temperature.

* 1. **Khám phá tập dữ liệu Tín dụng Đức :**

1.

- Nội dung phần ghi chú trong credit-g.arff nói về :

+ Tiêu đề tập dữ liệu : German Credit data

+ Thông tin nguồn:

Text

Description automatically generated

+Số lượng mẫu : 1000

+Số lượng thuộc tính : 20 (7 numerical, 13 categorical)

+ Các mô tả về các thuộc tính.

2. Tên thuộc tính lớp là class. Phân bố lệch về good.

Chart, bar chart

Description automatically generated