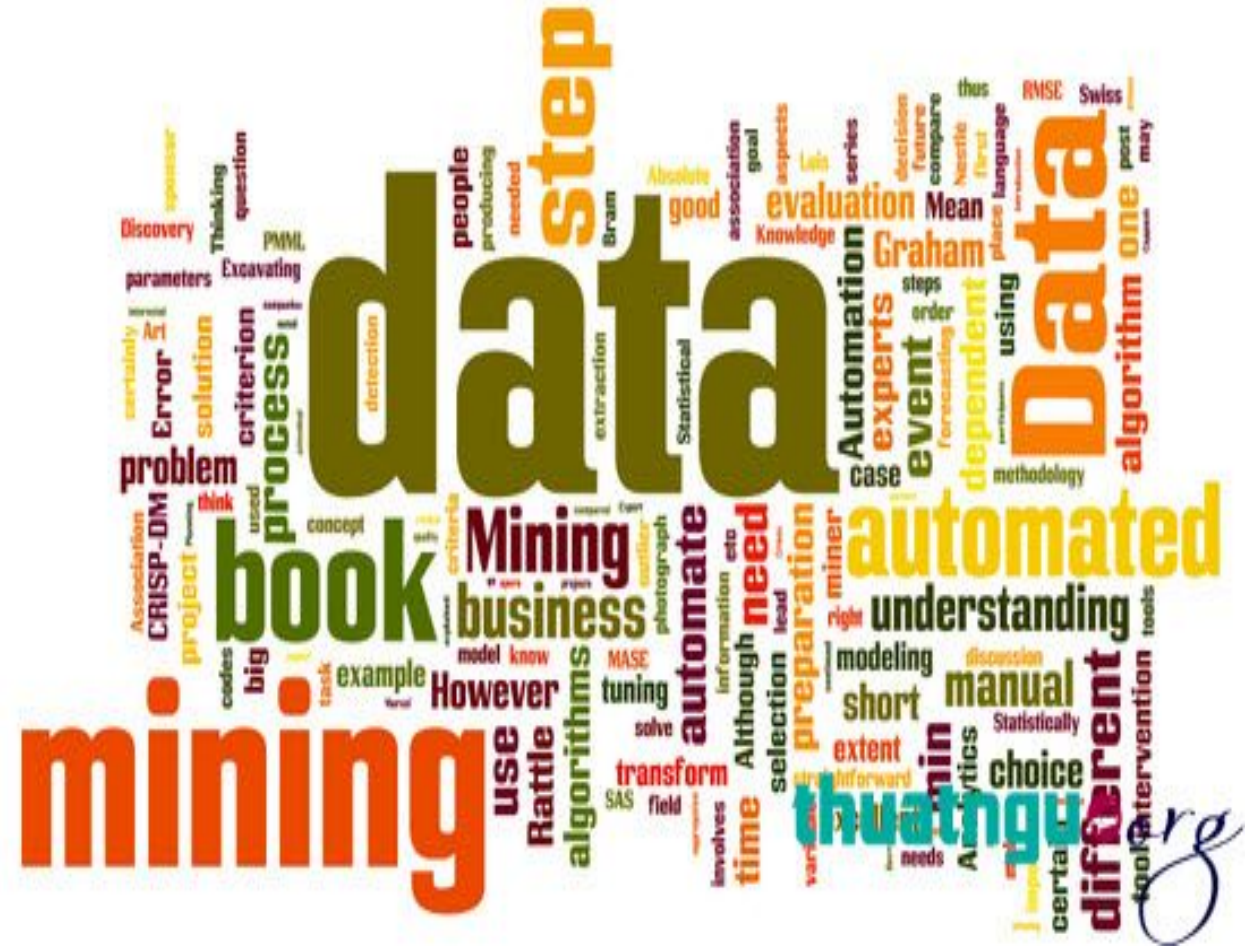
KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

HƯỚNG DẪN SỬ DỤNG CHƯƠNG TRÌNH TỰ CÀI ĐẶT CỦA LAB01 – Preprocessing

Hướng dẫn sử dụng các hàm đã cài đặt theo yêu cầu trong bài tập để ứng dụng thủ tục tiền xử lý dữ liệu đơn giản.



Yêu cầu của đề

Cài đặt chương trình tiền xử lý hai tập dữ liệu heart-h.arff và heart-c.arff với các chức năng chính như sau:

a. Đọc hai tập dữ liệu heart-h.arff và heart-c.arff. Cho phép người dùng truy vấn một dòng dữ liệu bất kỳ trong các tập dữ liệu này bằng cách nhập chỉ mục của dòng dữ liệu (tính từ 0).

Ví dụ, người dùng chọn tập tin heart-h.arff và chọn chỉ mục 0 thì xuất ra:

28,male,atyp_angina,130,132,f,left_vent_hyper,185,no,0,?,?,?,'<50'

b. Tích hợp tự động hai tập dữ liệu heart-h.arff và heart-c.arff thành tập dữ liệu mới heart-integration-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết. Đối chiếu với tập tin heart-integration.arff.

c. Làm sạch tự động tập dữ liệu heart-integration-auto.arff thành tập dữ liệu mới heart-cleaned-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết. Đối chiếu với tập tin heart-cleaned.arff.

d. Chuẩn hóa tự động tập dữ liệu heart-cleaned-auto.arff thành tập dữ liệu mới heart-normal-auto.arff theo các kỹ thuật đã lựa chọn trong phần báo cáo viết. Đối chiếu với tập tin heart-normal.arff.

e. Lấy mẫu tự động theo phương pháp Simple Random Sample Without Replacement và Simple Random Sample Without Replacement. Phát sinh ra hai tập dữ liệu mới tương ứng là heart-srswr.arff và heart-srsowr.arff, mỗi tập tin chứa 50% dữ liệu cũ từ tập dữ liệu heart-normal.arff.

Tổ chức các lớp (class)

Package: main

BaseApplication: chứa các hàm đọc ghi dữ liệu ra file và lấy dữ liệu từ file.

Main: chứa hàm main thực thi chương trình.

CauA, CauB, CauC, CauD, CauE: chứa các hàm xử lí theo theo cầu từ đề.

Package: model

Attribute: cấu trúc của 1 thuộc tính cần phải có.

Danh sách hàm cần thiết và cách gọi trong các lớp (class) theo yêu cầu đề

CauA

CauA.timKiemChiMuc(): hàm tìm kiếm theo chỉ mục.

- **Đầu vào:** không có
- **Đầu ra:** không có

Mô tả: Khi vào hàm sẽ có câu thông báo và yêu cầu người dùng **chọn tên tập tin** cần truy vấn dữ liệu, theo yêu cầu đề thì ta có 2 tập dữ liệu để truy vấn trong file là ***heart-h.arff*** và ***heart-c.arff***. Sau khi chọn 1 trong 2 tập tin truy vấn xong ta tiếp tục nhập số thứ tự dữ liệu cần truy vấn (bắt đầu từ 0 đến dòng dữ liệu cuối cùng). **Lưu ý:** nếu nhập dòng dữ liệu cần truy vấn bằng 0 hoặc số thự tự dòng dữ liệu đó vượt quá số dữ liệu trong tập dữ liệu thì màn hình sẽ không hiện dòng dữ liệu mà người dùng muốn truy vấn.

CauB

void CauB.combineFiles(File file1, File file2): hàm tích hợp 2 dữ liệu vào 1 file (tên file theo yêu cầu đề cố định).

- **Đầu vào:** File file1 - cấu trúc file đầu tiên cần tích hợp.

  File file2 - cấu trúc file thứ 2 cần tích hợp

- **Đầu ra:** không có.

Mô tả: hàm sẽ lấy tên cột dữ liệu trong 2 tập tin tích hợp lại với nhau. Sau đó, lấy tập dữ liệu của 2 tập tin tích hợp lại với nhau dựa vào cột dữ liệu của 2 tập tin. Cuối cùng là xuất

ra tập tin chứa kết quả tích hợp 2 tập dữ liệu của 2 tập tin vào 1 tập tin (tên là **heart-integration-auto.arff**).

CauC

void CauC.cleanupFiles(): hàm làm sạch dữ liệu, lấy dữ liệu từ tập tin **heart-integration-auto.arff** xuất dữ liệu sau khi làm sạch ra tập tin **heart-cleaned-auto.arff.**

- **Đầu vào:** không có
- **Đầu ra:** không có

CauD

void CauD.standardlized(): hàm chuẩn hóa tự động tập dữ liệu **heart-cleaned-auto.arff** thành tập dữ liệu mới **heart-normal-auto.arff.**

- **Đầu vào:** không có
- **Đầu ra:** không có

CauE

boolean CauE.SimpleRandSampleWithoutReplacement(Float percent): hàm lấy mẫu tự động theo phương pháp Simple Random Sample Without Replacement. Dữ liệu lấy từ file **heart-normal.arff** cố định. Sinh ra tập dữ liệu mới tên cố định **heart-srsowr.arff**.

- **Đầu vào: Float percent** - phần trăm dữ liệu cũ cần lấy.
- **Đầu ra: boolean -** khi dữ liệu xử lí thành công (xuất ra file thành công) hàm trả về true, khi có lỗi xảy ra thì hàm sẽ trả về false (nếu trường hợp lỗi ở xuất dữ liệu ra file thì hàm vẫn tạo ra file nhưng không đủ dữ liệu).

boolean CauE.SimpleRandSampleWithReplacement(Float percent): hàm lấy mẫu tự động theo phương pháp Simple Random Sample With Replacement. Dữ liệu lấy từ file **heart-normal.arff** cố định. Sinh ra tập dữ liệu mới tên cố định **heart-srswr.arff**.

- **Đầu vào: Float percent** - phần trăm dữ liệu cũ cần lấy.

- **Đầu ra: boolean** - khi dữ liệu xử lí thành công (xuất ra file thành công) hàm trả về true, khi có lỗi xảy ra thì hàm sẽ trả về false (nếu trường hợp lỗi ở xuất dữ liệu ra file thì hàm vẫn tạo ra file nhưng không đủ dữ liệu).