Trường Đại học Khoa học Tự nhiên – ĐHQG TP.HCM
Khoa Công nghệ thông tin
Bộ Môn Khoa học máy tính

BÁO CÁO ĐỒ ÁN
CLUSTERING & CLASSIFICATION
Khai thác dữ liệu & Ứng dụng

Đặng Minh Thọ - MSSV: 18120579
Phạm Đình Thục - MSSV: 18120584

Thành phố Hồ Chí Minh, tháng 12, năm 2020

I. PHÂN CÔNG CÔNG VIỆC

MSSV	Công việc	Tỷ lệ hoàn thành
18120579	Phân lớp dữ liệu bằng Weka Explorer	50%
	Thuật toán K medoids	
18120584	Phân lớp dữ liệu bằng Weka Experimenter	50%
	Thuật toán K means	

TỔNG: 100%

II. PHÂN LỚP DỮ LIỆU BẰNG WEKA

III. PHÂN LỚP DỮ LIỆU BẰNG WEKA EXPERIMENTER

Tiền xử lý:

- Xoá các thuộc tính có tỷ lệ thiếu dữ liệu lớn hơn hoặc bằng 50%
- Xoá các dữ liệu dạng Identification
- Điền giá trị thiếu tại các cột có kiểu dữ liệu dạng số bằng giá trị trung bình
- Điều giá trị thiếu cho các cột có kiểu dữ liệu định danh bằng giá trị mode

IV. ĐÁNH GIÁ

- Phương pháp phân lớp nào thường cho kết quả cao nhất?
  ➔ J48 có tỉ lệ cao nhất (97.98% ở thực nghiệm D) và Id3 có tỉ lệ cao nhất (100% ở thực nghiệm A-C)

- Phương pháp nào không thực hiện tốt và tại sao?
  ➔ Id3 có tỉ lệ thấp nhất 48.2201% (Percentage split với 66%)

- Tại sao ta sử dụng phiên bản đã rời rạc hóa của tập dữ liệu nếu tập dữ liệu đã được rời rạc hóa?
  ➔ Vì các dữ liệu chứa giá trị rời rạc đóng vai trò quan trọng trong việc biểu diễn tri thức vì chúng dễ dàng được xử lý cũng như thể hiện tri thức trực quan hơn

- Việc rời rạc hóa và cách rời rạc hóa có ảnh hưởng đến kết quả phân lớp hay không, nếu có thì ảnh hưởng thể nào?
  ➔ Không ảnh hưởng nhiều đến độ chính xác của kết quả đối với các phân lớp khác nhưng nếu không rời rạc hóa thì không thực hiện được phân lớp Id3 do chứa thuộc tính liên tục

- Chiến lược nào trong ba chiến lược đánh giá đã đánh giá quá cao (overestimate) độ chính xác và tại sao?
  ➔ Using training test vì nó sử dụng tập test từ tập dữ liệu cũ nên kết quả đánh giá cao độ chính xác

- Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao?
  ➔ Percentage split với 66% vì chỉ định tỷ lệ phân chia tập dữ liệu đối với việc đánh giá là 66% (dùng 33% làm tập dữ liệu test và 66% làm tập dữ liệu để train) mà không sử toàn bộ tập dữ liệu để train nên kết quả đánh giá thấp độ chính xác