TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HCM KHOA CÔNG NGHỆ THÔNG TIN

DATA MINING

BÁO CÁO LAB3

SINH VIÊN THỰC HIỆN:

1712919 Lê Văn Vũ

GV LÝ THUYẾT/ HD THỰC HÀNH:

Thầy Lê Hoài Bắc

Thầy Nguyễn Ngọc Đức

Thầy **Dương Nguyễn Thái Bảo**

Thầy **Hoàng Xuân Trường**

• Phương pháp phân lớp nào thường cho kết quả cao nhất?

PP phân lớp thường cho kết quả cao nhất là **Use training set**.

• Phương pháp nào không thực hiện tốt và tại sao?

PP phân lớp **Percentage split** cho kết qủa không tốt, vì khi tách tập dữ liệu theo tỷ lệ (thông số) không phù hợp sẽ làm mất đi tính liên quan của dữ liệu, nhiều trường hợp xấu có thể làm mất/sai lệch thông tin của một lớp phân loại.

Hơn nữa, với dữ liệu hawk này chỉ có 907 mẫu, cộng thêm số lượng dữ liệu bị thiếu quá nhiều nên pp này cho kết quả không đồng đều.

• 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 rời rạc hóa các thuộc tính số để các thuộc tính được mô tả đúng như ý nghĩa của nó.

• 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?

Qua quá trình thực nghiệm bằng cách rời rạc hóa các thuộc tình không phải là lớp (B và C) thì nhận thấy việc rời rạc và cách rời rạc hầu như **không ảnh hưởng** đến kết quả phân lớp.

• Chiến lược nào trong ba chiến lược đánh giá đã đánh giá quá cao (overestimate) độ chính xác và tại sao?

Hai thuật toán NaiveBayesSimple và J48 cho kết quả rất tốt, nhưng suy cho cùng thì J48 tốt hơn cả.

Bởi lẽ, các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu Qualititive value (ordinal, Binary, nominal).

- Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao?

Khi đánh giá bằng thuật toán Id3 ta nhận thấy kết quả chênh lệch 'rất lớn' giữa các loại thwucj nghiệm.

Thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số (quantitative value).