Đại học Quốc gia TP.HCM
Trường Đại học Khoa học Tự nhiên

CLASSIFICATION & CLUSTERING

Khai thác dữ liệu & Ứng dụng

Nguyễn Bảo Long - MSSV: 18120201
Huỳnh Long Nam - MSSV: 18120212

TP Hồ Chí Minh, ngày 02/12/2020

Contents

1 Thông tin chung

1. Link GitHub: `https://github.com/baolongnguyenmac/DataMining-Lab3`
2. Thông tin thành viên nhóm

STT	Họ tên	MSSV	Email
1	Nguyễn Bảo Long	18120201	18120201@student.hcmus.edu.vn
2	Huỳnh Nam Long	18120212	18120212@student.hcmus.edu.vn

Table 1: Bảng thông tin thành viên nhóm

3. Tỷ lệ tham gia công việc

STT	Họ tên	Công việc	Tỷ lệ hoàn thành
1	18120201	Tiền xử lý dữ liệu	50%
2		Phân lớp dữ liệu bằng Weka Explorer	
3		Cài đặt và kiểm thử thuật toán K-Means	
4		Trình bày báo cáo	
5	18120212	Phân lớp dữ liệu bằng Experimenter	50%
6		Đánh giá phương pháp phân lớp	
7		Cài đặt và kiểm thử thuật toán K-Medoids	

Table 2: Bảng phân chia công việc

2 Ý tưởng tiền xử lý dữ liệu

- Chi tiết quá trình tiền xử lý dữ liệu được trình bày trong file 'Preprocess.ipynb'
- Ý tưởng chung
  - 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

3 Đánh giá phương pháp phân lớp

- Phương pháp phân lớp nào thường cho kết quả cao nhất?
  → Theo file "Result.xlsx", khi sắp xếp giảm dần cột "Tỷ lệ mẫu được phân đúng", ta thấy ID3 là thuật toán cho kết quả phân lớp đúng cao nhất (100%), kế đến là J48 (99.45%). Nhưng xét tổng quan, với cùng một chiến lược đánh giá, J48 cho kết quả cao và ổn định

PHÂN LỚP DỮ LIỆU BẰNG WEKA EXPLORER

STT	Loại thực nghiệ	Tên tệp tin dữ liệu đầu và	Phương pháp phân lớp	Chiến lược đánh giá	Tỷ lệ mẫu được phân đún
13	B	afterPreprocess.csv	ID3	Using Training Test	100%
22	C	afterPreprocess.csv	ID3	Using Training Test	100%
7	A	afterPreprocess.csv	J48	Using Training Test	99.45%

Figure 1: Thống kê tỷ lệ mẫu được phân lớp đúng ứng với các thuật toán

- Phương pháp nào không thực hiện tốt và tại sao?
  → ID3 fit với bộ training test nên độ chính xác khi đánh giá sử dụng bộ training test sẽ cho 100%
  → Phương pháp ID3 không thực hiện tốt vì nhược điểm của ID3 là sẽ quá khớp với bộ dữ liệu train, nên khi gặp những bộ dữ liệu bị nhiễu sẽ làm cây quyết định cồng kềnh và tỉ lệ dự đoán trên tập test thấp

- 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 giúp giảm miền giá trị của các thuộc tính, khiến cây đơn giản hơn, thuật toán chạy nhanh hơn đồng thời cũng khử nhiễu

- 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?
  →Việc rời rạc hóa có ảnh hưởng đến kết quả phân lớp
  →Nếu rời rạc hóa quá ít, thì luật rút ra sẽ có độ tin cậy thấp, hiệu quả phân lớp không cao
  →Nếu rời rạc hóa quá nhiều, luật rút ra sẽ quá khớp với bộ dữ liệu train, cây phân lớp sẽ lớn, thuật toán chạy lâu

- Chiến lược nào trong ba chiến lược đánh giá đã đánh giá quá cao (overestimate) độ chính xác và tại sao?
  → Chiến lược đánh giá Using Traning Test đã đánh giá quá cao độ chính xác.
  → Vì các thuật toán chạy trên bộ dữ liệu huấn luyện nên sẽ cho ra mô hình khớp với bộ dữ liệu huấn luyện, nên độ chính xác sẽ cao trên tập huấn luyện

- Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao?
  → Chiến lược đánh giá Percentage split với 66% đã đánh giá thấp độ chính xác so với 2 chiến lược đánh giá còn lại,
  → Vì mô hình chỉ khớp với với tập train, mà không chắc rằng tập train với tập test tương đồng nhau, nên mô hình sẽ không cho kết quả tốt trên tập test