**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**
KHOA HỆ THỐNG THÔNG TIN

**TÀI LIỆU HƯỚNG DẪN THỰC HÀNH**
Môn: Khai thác dữ liệu

# BÀI THỰC HÀNH SỐ 3

## I. Mục tiêu

1. Hiểu rõ các lý thuyết và kỹ thuật của tập thô.
2. Áp dụng các kỹ thuật trong tập thô vào bài toán lựa chọn thuộc tính và khai phá luật phân lớp.

## II. Thời gian

1. Thực hành: 4 tiết
2. Bài tập làm thêm: 8 tiết

## III. Thực hành

1. (Cơ bản) Cho cơ sở dữ liệu sau:

| **Đối tượng** | **Các thuộc tính điều kiện** | | | **Thuộc tính quyết định** |
|---|---|---|---|---|
| | **C** | | | **D** |
| **U** | **Máy** | **Kích thước** | **Màu** | **d** |
| x1 | Prôban | Gọn | Đen | Tốt |
| x2 | Dầu | Trung bình | Vàng | Xấu |
| x3 | Dầu | Lớn | Trắng | Xấu |
| x4 | Dầu | Trung bình | Đỏ | Xấu |
| x5 | Xăng | Gọn | Đen | Tốt |
| x6 | Xăng | Trung bình | Bạc | Tốt |
| x7 | Xăng | Lớn | Trắng | Xấu |
| x8 | Xăng | Gọn | Bạc | Xấu |

Tìm quan hệ bất khả phân biệt IND(A) và các lớp tương đương đối với các tập thuộc tính

a. A = Q {Toàn bộ các thuộc tính}
b. A = {Máy}
c. A = {Kích thước}
d. A = {Màu}
e. A = {d}

f. A = {Máy, Kích thước}

g. A = {Máy, Màu}

h. A = {Kích thước, Màu}

i. A = {Màu, d}

j. A = {Máy, Kích thước, Màu}

Tìm xấp xỉ trên, xấp xỉ dưới và đường biên trên tập thuộc tính A

k. A = {Máy, Màu}, X = {x1, x2, x3, x4}

l. A = {Máy, Kích thước}, X = {x1, x2, x3, x6}

m. A = {Máy, Kích thước, Màu}, X = {x1, x2, x4, x6}

Tìm luật quyết định với

n. A = {Máy, Kích thước, Màu}

o. A = {Máy, Kích thước}

p. A = {Kích thước, Màu}

2. (Cơ bản) Cho cơ sở dữ liệu về máy tính như sau:

| **Đối tượng** | **Các thuộc tính điều kiện** | | | **Thuộc tính quyết định** |
|---|---|---|---|---|
| **U** | **C** | | | **D** |
| **PC** | **MONITOR** | **OS** | **CPU** | **d** |
| x1 | Màu | Linux | arm | Tốt |
| x2 | Màu | Windows | x86 | Xuất sắc |
| x3 | Trắng đen | Linux | x86 | Xấu |
| x4 | Màu | Windows | arm | Xuất sắc |
| x5 | Màu | Linux | x64 | Xấu |
| x6 | Trắng đen | Windows | arm | Tốt |
| x7 | Trắng đen | Windows | x86 | Tốt |
| x8 | Màu | Windows | x64 | Tốt |

Tìm quan hệ bất khả phân biệt IND(A) và các lớp tương đương đối với các tập thuộc tính

a. A = Q {Toàn bộ các thuộc tính}

b. A = {MONITOR}

c. A = {OS}

d. A = {CPU}

e. A = {d}

f. A = {MONITOR, OS}

g. A = {MONITOR, CPU}

h. A = {OS, CPU}

i. A = {OS, d}

j. A = {MONITOR, OS, CPU}

Tìm xấp xỉ trên, xấp xỉ dưới và đường biên trên tập thuộc tính A

k. A = {MONITOR}, X = {x1, x2, x3, x4}

l. A = {MONITOR, OS}, X = {x1, x2, x3, x6}

m. A = { MONITOR, OS, CPU}, X = {x1, x2, x4, x6}

Tìm luật quyết định với

n. A = {MONITOR, OS, CPU}

o. A = {MONITOR, CPU}

p. A = {OS, CPU}

3. (Cơ bản) Cho CSDL về Theo dõi Khuyến mãi của một công ty thời trang, cho bởi bảng sau. Ghi chú:

- Thuộc tính Mua hàng (MH) là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài

| | **Giới tính (GT)** | **Tuổi (T)** | **Lần mua gần nhất (LMGN)** | **Khuyến mãi (KM)** | **Mua hàng (MH)** |
|---|---|---|---|---|---|
| 1 | Nữ | 20..25 | < 1 tháng | Upto 70% | Có |
| 2 | Nam | 20..25 | 1..3 tháng | Upto 70% | Không |
| 3 | Nữ | 26..30 | >3 tháng | 1+1 | Có |
| 4 | Nữ | >30 | 1..3 tháng | 30% | Có |
| 5 | Nam | 26..30 | >3 tháng | 30% | Có |
| 6 | Nữ | 26..30 | >3 tháng | 1+1 | Không |
| 7 | Nữ | >30 | >3 tháng | 1+1 | Không |
| 8 | Nam | 26..30 | < 1 tháng | 30% | Không |
| 9 | Nữ | >30 | 1..3 tháng | Upto 70% | Không |
| 10 | Nữ | 26..30 | < 1 tháng | Upto 70% | Có |

Sử dụng tập thô tính: *xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ.*

a. Với B={Lần mua gần nhất, Khuyến mãi}, X={1, 3, 4, 5, 10} (tập các mẫu có giá trị Mua Hàng = “Có”).

b. Với B={Tuổi, Lần mua gần nhất}, X={2, 6, 7, 8, 9} (tập các mẫu có giá trị Mua Hàng = "Không").

4. (Cơ bản) Cho CSDL về *Xét nghiệm SARS-CoV-2* của một khu cách ly, cho bởi bảng sau. Ghi chú:
   - Thuộc tính *Kết quả (KQ)* là thuộc tính quyết định.
   - Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài

| | **Cư trú (CT)** | **Tuổi (T)** | **Trình trạng hô hấp (TTHH)** | **Thân nhiệt (TN)** | **Kết quả (KQ)** |
|---|---|---|---|---|---|
| 1 | Trong nước | <25 | Khó thở | <37 | Âm tính |
| 2 | Nước ngoài | >55 | Khó thở | <37 | Âm tính |
| 3 | Trong nước | >55 | Khó thở | >38 | Dương tính |
| 4 | Nước ngoài | 25..55 | Hụt hơi | 37..38 | Dương tính |
| 5 | Nước ngoài | <25 | Bình thường | 37..38 | Dương tính |
| 6 | Trong nước | 25..55 | Bình thường | <37 | Âm tính |
| 7 | Trong nước | >55 | Hụt hơi | >38 | Âm tính |
| 8 | Trong nước | <25 | Hụt hơi | 37..38 | Dương tính |
| 9 | Nước ngoài | >55 | Khó thở | <37 | Dương tính |
| 10 | Nước ngoài | 25..55 | Bình thường | <37 | Âm tính |

   Cho B={Tình trạng hô hấp, Thân nhiệt}, X={3, 4, 5, 8, 9} (tập các mẫu có giá trị Kết quả = "Dương tính"). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ.

## IV. Bài tập thêm

1. Sử dụng ngôn ngữ lập trình Python, cài đặt các giải thuật tìm xấp xỉ trên, dưới, vùng biên, vùng ngoài.
2. Sử dụng ngôn ngữ lập trình Python, cài đặt giải thuật tìm các rút gọn và tập lõi.

## V. Tài liệu tham khảo


1. **Jiawei Han, Micheline Kamber, and Jian Pei,** *Data Mining Concepts and Techniques,* 3 edition, Morgan Kaufmann Publishers, 2011.
2. **Graham J. Williams, Simeon J. Simoff**, *"Data Mining: Theory, Methodology, Techniques, and Applications"*, Springer-Verlag, 2006.