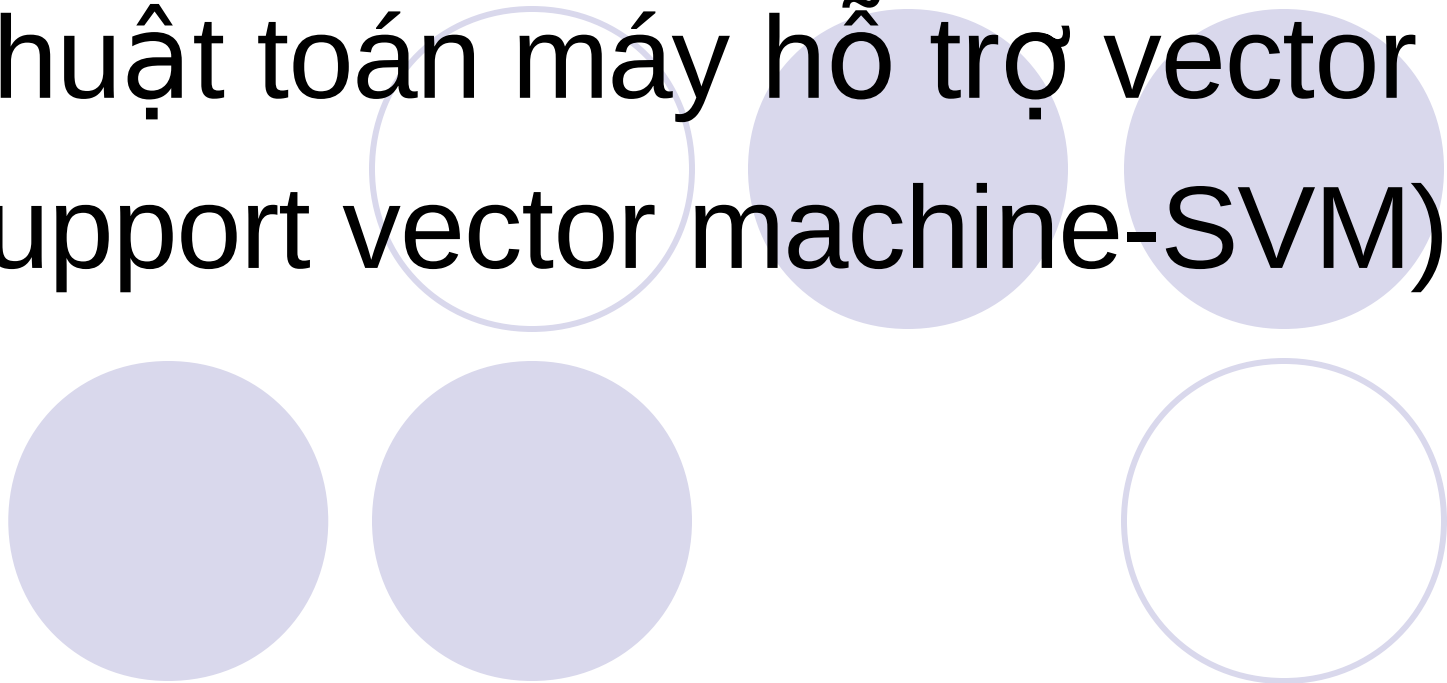


Thuật toán máy hỗ trợ vector (support vector machine-SVM)

The slide features several decorative circles. There are three solid light purple circles in the upper right area, one solid light purple circle in the lower left, and two hollow light purple circles in the lower right.

Người hướng dẫn:CN.Vũ Tiến Thành

Sinh viên :Đinh thị Hương

Lớp : k52CA



Nội dung chính

I. Phát biểu bài toán.

1. Trình bày tóm tắt về phân lớp dữ liệu
2. Tại sao lại sử dụng thuật toán SVM để phân lớp dữ liệu?

II. Thuật toán máy hỗ trợ vector(SVM)

1. Định nghĩa
2. Mục đích
3. Mô hình giải thuật
4. Các câu hỏi
5. Giải quyết
6. Transductive support vector machine(TSVM)

III. Ứng dụng của SVM

I. Phát biểu bài toán:

1. Trình bày tóm tắt về phân lớp dữ liệu

- **Phân lớp dữ liệu** là một kỹ thuật trong khai phá dữ liệu được sử dụng rộng rãi nhất và được nghiên cứu mở rộng hiện nay.
- **Mục đích:** Để dự đoán những nhãn phân lớp cho các bộ dữ liệu hoặc mẫu mới.
- **Đầu vào:** Một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- **Đầu ra:** Bộ phân lớp dựa trên tập huấn luyện, hoặc những nhãn phân lớp



- Kỹ thuật phân lớp dữ liệu được tiến hành bao gồm 2 bước:

Bước 1:

Xây dựng mô hình từ tập huấn luyện

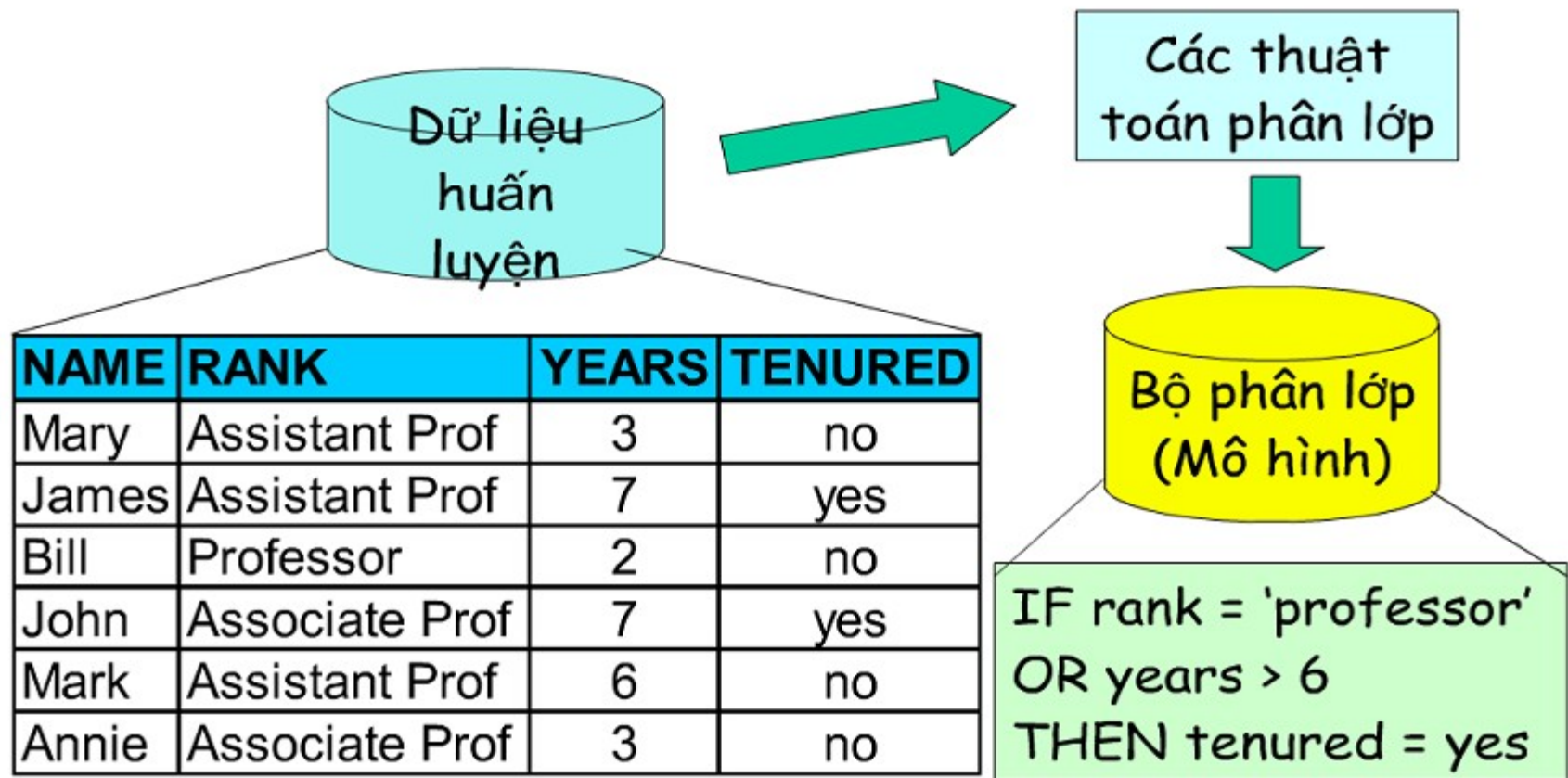
Bước 2:

Sử dụng mô hình – kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới.

Bước 1. Xây dựng mô hình

- Mỗi bộ/mẫu dữ liệu được phân vào một lớp được xác định trước.
- Lớp của một bộ/mẫu dữ liệu được xác định bởi thuộc tính gán nhãn lớp
 - Tập các bộ/mẫu dữ liệu huấn luyện-tập huấn luyện – được dùng để xây dựng mô hình.
- Mô hình được biểu diễn bởi các luật phân lớp, các cây quyết định hoặc các công thức toán học.

Ví dụ: xây dựng mô hình



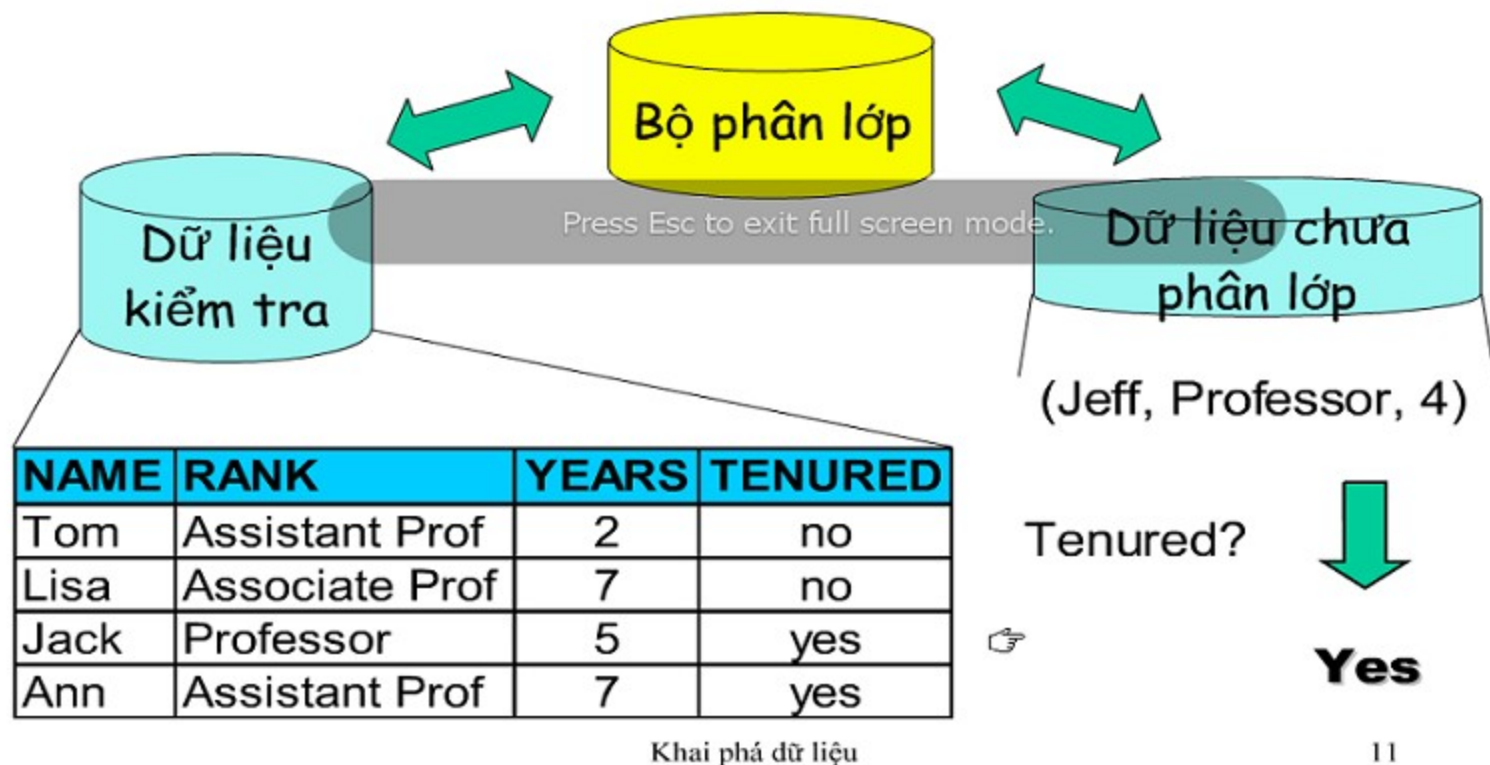
Khai phá dữ liệu

10

Bước 2: Sử dụng mô hình

- Phân lớp cho những đối tượng mới hoặc chưa được phân lớp
- Đánh giá độ chính xác của mô hình
 - + Lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình.
 - + Tỷ lệ chính xác bằng phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra.

Ví dụ: sử dụng mô hình





- Các thuật toán phân lớp dữ liệu phổ biến:

- + Thuật toán cây quyết định
- + Thuật toán SVM
- + Thuật toán phân lớp Bayes
- + Thuật toán phân lớp K người láng giềng gần nhất.

2. Tại sao lại sử dụng thuật toán SVM trong phân lớp dữ liệu?

- SVM rất hiệu quả để giải quyết bài toán dữ liệu có số chiều lớn (ảnh của dữ liệu biểu diễn gene, protein, tế bào)
- SVM giải quyết vấn đề overfitting rất tốt (dữ liệu có nhiều và tách rời nhóm hoặc dữ liệu huấn luyện quá ít)
- Là phương pháp phân lớp nhanh
- Có hiệu suất tổng hợp tốt và hiệu suất tính toán cao.