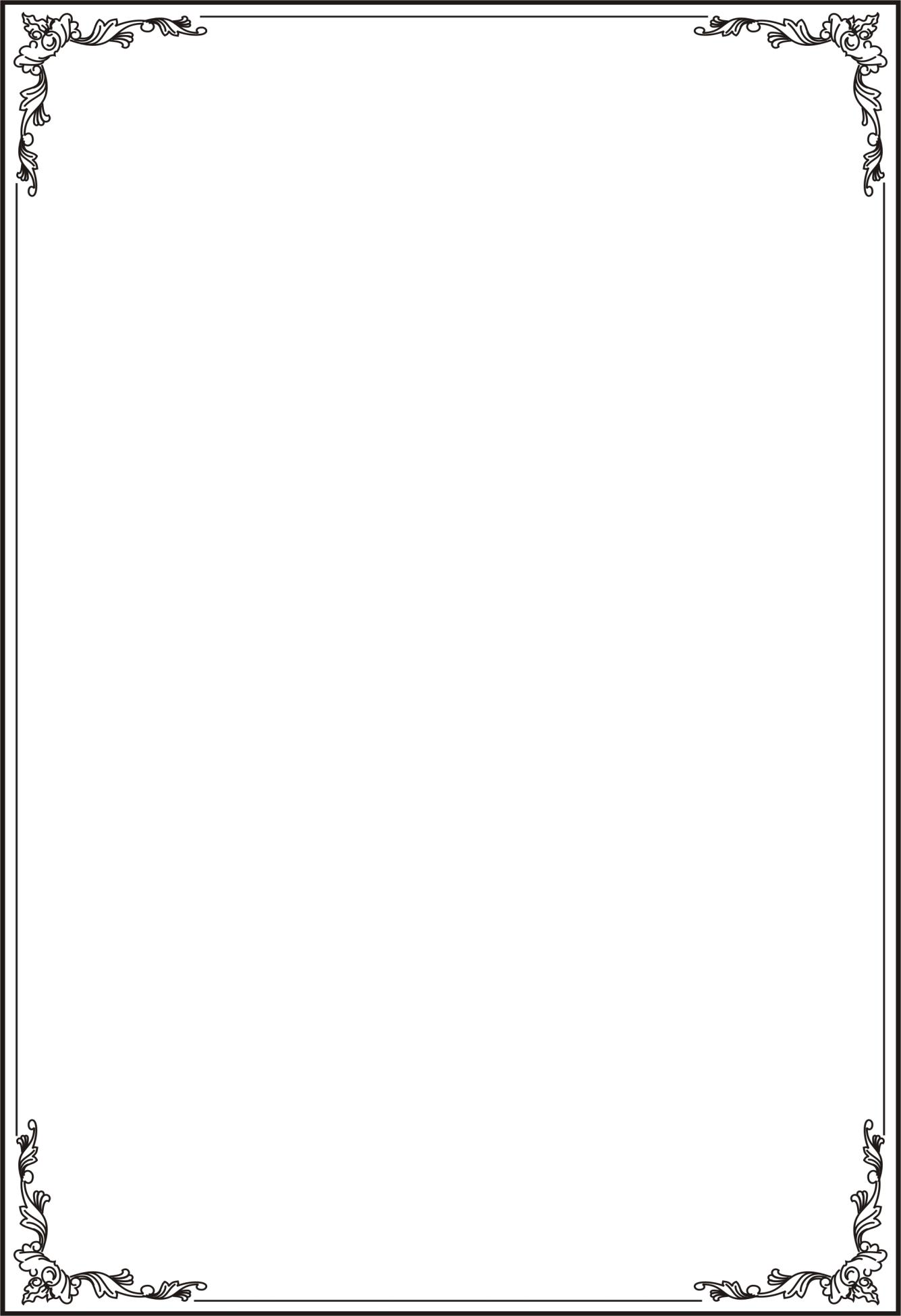
****

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**BÁO CÁO ĐỒ ÁN DATAMINING  
*Đề tài: Phân lớp dáng bộ người dùng***

**Lớp: CS313.I21.KHTN**

**GVLT: Nguyễn Hoàng Tú Anh**

**GV HDTH: Nguyễn Thị Anh Thư**

**Nhóm thực hiện:**

1. **Nguyễn Quốc Danh – 15520092**
2. **Nguyễn Minh Dũng – 15520138**

**Tp.Hồ Chí Minh, ngày 5 tháng 5 năm 2018**

**NHẬN XÉT**

Menu

[1 MÔ TẢ TÓM TẮT VỀ TẬP DỮ LIỆU 1](#_Toc513343780)

[2 CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU 1](#_Toc513343781)

[3 PHƯƠNG PHÁP SỬ DỤNG TRONG CHƯƠNG TRÌNH 1](#_Toc513343782)

[3.1 Tên phương pháp 1](#_Toc513343783)

[3.2 Các đặc trưng chính của phương pháp 1](#_Toc513343784)

[3.3 Thuật toán 1](#_Toc513343785)

[4 CÁC THỰC NGHIỆM ĐÃ TIẾN HÀNH 1](#_Toc513343786)

[4.1 Thực nghiệm 1 1](#_Toc513343787)

# 

# MÔ TẢ TÓM TẮT VỀ TẬP DỮ LIỆU

- Tập dữ liệu được thu thập dùng để

- Có tất cả 165633 mẫu trong tập dữ liệu

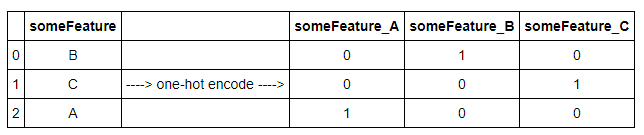
- Mỗi mẫu có 18 thuộc tính

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Công dụng | Trung bình/số giá trị phân biệt | Phương sai/Số giá trị duy nhất | Số mẫu bị thiếu |
| 1 | user | nominal | Tên người dùng | 4 | 0 | 0(0%) |
| 2 | gender | nominal | Giới tính | 2 | 0 | 0(0%) |
| 3 | age | numeric | Tuổi | 38.265 | 13.184 | 0(0%) |
| 4 | how\_tall\_in\_meters | numeric | Chiều cao | 1.64 | 0.053 | 0(0%) |
| 5 | weight | numeric | Cân nặng | 70.819 | 11.297 | 0(0%) |
| 6 | body\_mass\_index | numeric | Chỉ số khối cơ thể | 26.189 | 2.996 | 0(0%) |
| 7 | x1 | numeric | Giá trị x của gia tốc kế gắn trên eo | -6.649 | 11.616 | 0(0%) |
| 8 | y1 | numeric | Giá trị y của gia tốc kế gắn trên eo | 88.294 | 23.896 | 0(0%) |
| 9 | z1 | numeric | Giá trị z của gia tốc kế gắn trên eo | -93.165 | 39.409 | 0(0%) |
| 10 | x2 | numeric | Giá trị x của gia tốc kế gắn trên đùi trái | -87.828 | 169.435 | 0(0%) |
| 11 | y2 | numeric | Giá trị y của gia tốc kế gắn trên đùi trái | -52.065 | 205.16 | 0(0%) |
| 12 | z2 | numeric | Giá trị z của gia tốc kế gắn trên đùi trái | -175.055 | 192.817 | 0(0%) |
| 13 | x3 | numeric | Giá trị x của gia tốc kế gắn trên mắt cá phải | 17.424 | 52.635 | 0(0%) |
| 14 | y3 | numeric | Giá trị y của gia tốc kế gắn trên mắt cá phải | 104.517 | 54.156 | 0(0%) |
| 15 | z3 | numeric | Giá trị z của gia tốc kế gắn trên mắt cá phải | -93.882 | 45.39 | 0(0%) |
| 16 | x4 | numeric | Giá trị x của gia tốc kế gắn trên cánh tay phải | -167.641 | 38.311 | 0(0%) |
| 17 | y4 | numeric | Giá trị y của gia tốc kế gắn trên cánh tay phải | -92.625 | 19.969 | 0(0%) |
| 18 | z4 | numeric | Giá trị ­­­­­­­z của gia tốc kế gắn trên cánh tay phải | -159.651 | 13.221 | 0(0%) |

# CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

## **Chuyển dữ liệu nominal về dạng binary** (nominal to binary)

- Chuyển đổi dữ liệu nominal về binary theo dạng:

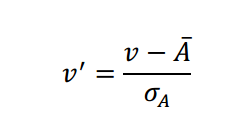
Ta có 2 thuộc tính nominal là “user” và “gender”

Thuộc tính “class” thì ta chuẩn hoá theo công thức:

{'sitting': 1, 'sittingdown': 2, 'standing': 3, 'standingup': 4, 'walking': 5}

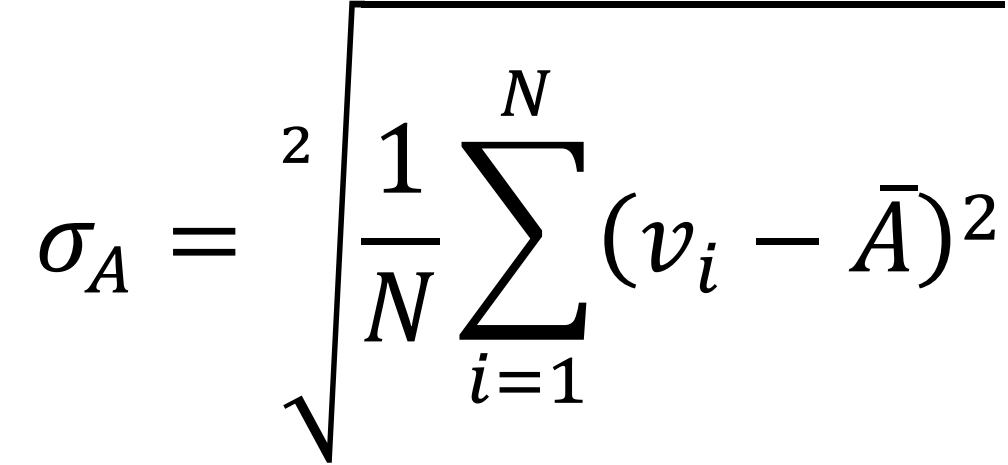
## Chuẩn hoá Z-Score

Các giá trị của thuộc tính sẽ được chuẩn hóa dựa trên giá trị trung bình và độ lệch chuẩn của thuộc tính đó.

Một giá trị v của thuộc tính A được chuẩn hóa theo Z-score sẽ được tính bằng công thức:

Trong đó:

- Ᾱ: là giá trị trung bình của thuộc tính. Tính bằng cách lấy trung bình cộng  
các giá trị thuộc tính A ở các mẫu trong tập dữ liệu.

- : độ lệch chuẩn của thuộc tính, tính theo công thức:

# PHƯƠNG PHÁP SỬ DỤNG TRONG CHƯƠNG TRÌNH

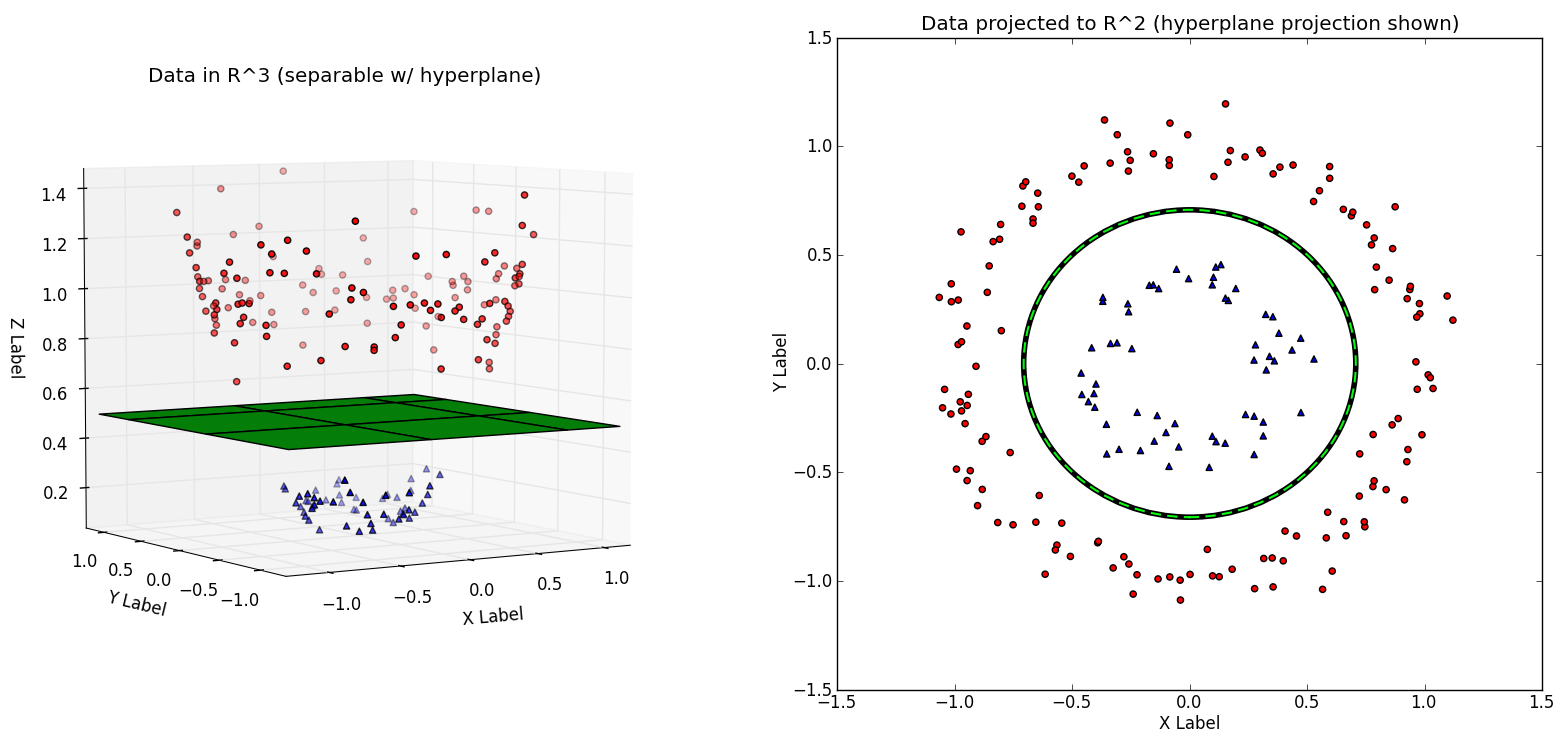
## Tên phương pháp

- Phương pháp sử dụng trong chương trình là Support Vector Machine (SVM)

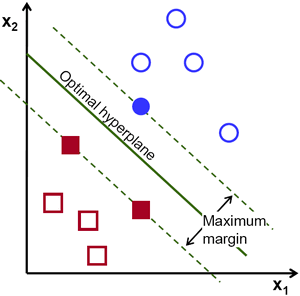
## Các đặc trưng chính của phương pháp

### Bài toán phân lớp và phương pháp Support Vector Machine

* Phân lớp là quá trình ‘nhóm’ các đối tượng ‘giống’ nhau vào một lớp dựa trên đặc trưng về dữ liệu của chúng. Khi xem xét một đối tượng chúng ta dựa vào một số đặc trưng hữu hạn của chúng. Hay nói cách khác ta sẽ biểu diễn một đối tượng trong một không gian hữu hạn chiều với mỗi chiều là một đặc trưng được lựa chọn.
* Phương pháp Support Vector Machine là phương pháp nhằm tìm ra siêu mặt phẳng có biên cực đại giữa lớp mẫu âm và lớp mẫu dương.

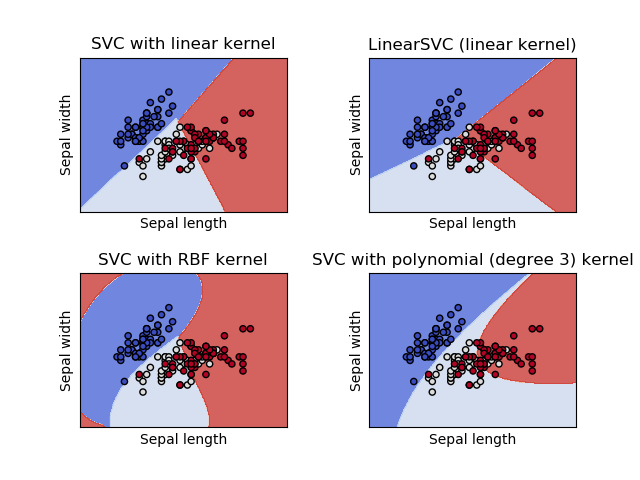


### Margin trong SVM:



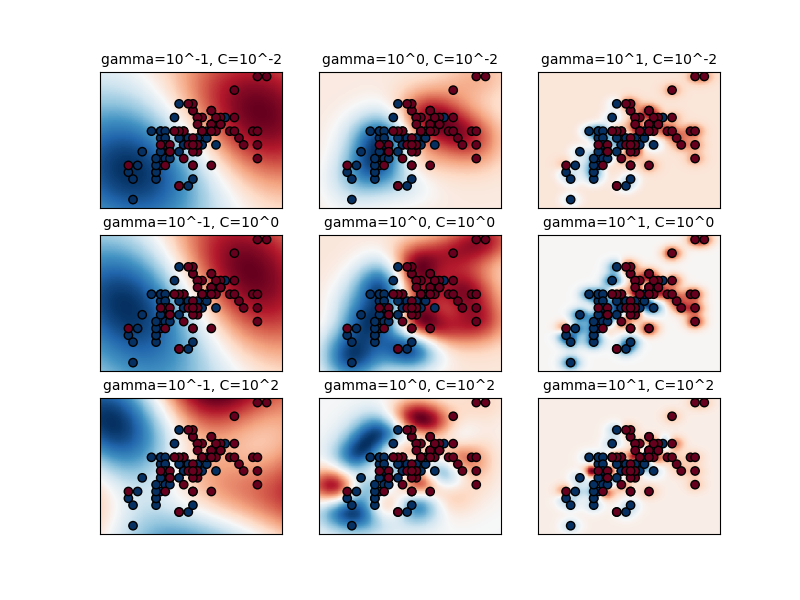
* Margin là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng của các phân lớp.
* Trong ví dụ trên siêu mặt phẳng chính là mặt phẳng mà khoảng cách của nó đến các điểm trong và vuông là lớn nhất. SVM luôn cố gắng cực đại hóa khoảng cách này để giảm thiểu việc phân lớp sai.

### Một số kernel trong SVM:



* Có thể thấy tùy theo dữ liệu ma ta chọn kernel phù hợp sao cho việc phân lớp được chính xác.

### Tham số C và Gamma trong SVM:



* Có thể thấy khi ta tăng tham số C lên thì vùng phân lớp thay đổi sẽ hạn chế điểm nhiễu.
* Tương tự khi tăng Gamma thì vùng phân lớp thu hẹp
* Chính vì thế ta phải chọn tham số cho C và Gamma phù hợp sao cho không quá lỏng và cũng không quá chặt để việc phân lớp được chính xác.

### Phân loại đa lớp

* Ý tưởng của bài toán phân lớp đa lớp là chuyển về bài toán phân lớp hai lớp bằng cách xây dựng nhiều bộ phân lớp hai lớp để giải quyết. Các chiến lược phân lớp đa lớp phổ biến này là One-against-One (OAO) và One-against-Rest (OAR).



Ví dụ phân lớp sử dụng chiến lược OAR và OAO

* Trong hình ta thấy chiến lược OAR (hình bên trái)- OAO (hình bên phải) phải xây dựng siêu phẳng để tách lớp đánh dấu “o” ra khỏi tất cả các lớp khác.

#### Chiến lược One-against-One (OAO).

* Trong chiến lược này ta sử dụng (n-1) bộ phân lớp đối với n lớp. Bài toán phân lớp n lớp được chuyển thành n bài toán phân lớp hai lớp.
* Nhược điểm của chiến lược OAR là ta phải xây dựng một siêu phẳng để tách một lớp ra khỏi các lớp còn lại, việc này đòi hỏi sự phức tạp và có thể không chính xác

#### Chiến lược One-against-Rest (OAR).

* Trong chiến lược này ta sử dụng n(n-1)/2 bộ phân lớp hai lớp được xây dựng bằng cách bắt cặp từng hai lớp một nên chiến lược này còn được gọi là pairwise và sử dụng phương pháp lựa chọn theo đa số để kết hợp các bộ phân lớp này để xác định được kết quả phân lớp cuối cùng. Số lượng các bộ phân lớp không bao giờ vượt quá n(n-1)/2.

# CÁC THỰC NGHIỆM ĐÃ TIẾN HÀNH

## Thực nghiệm

- Dữ liệu huấn luyện: 132506 mẫu

- Dữ liệu kiểm tra 1: 3313 mẫu

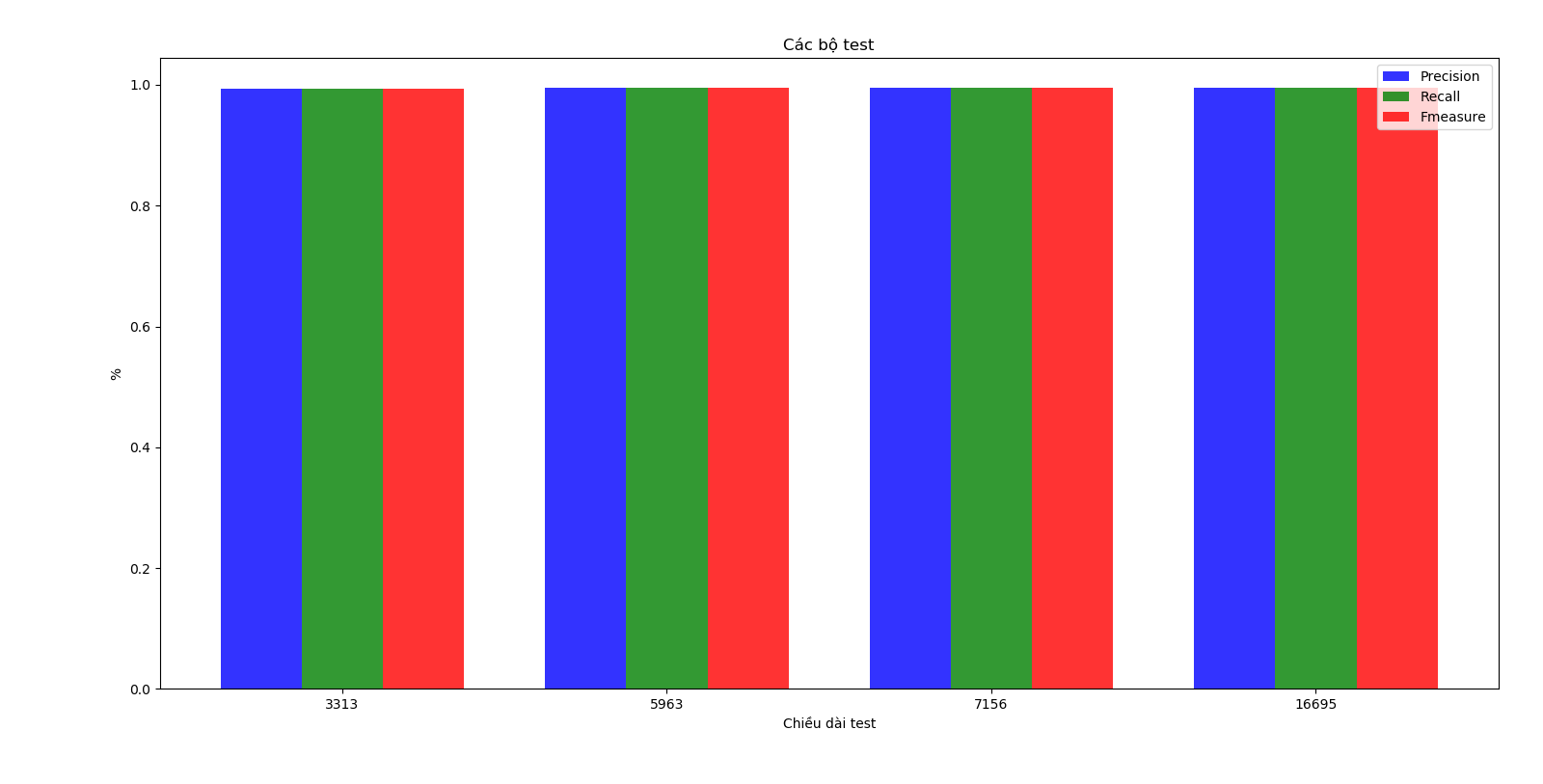
- Dữ liệu kiểm tra 2: 5963 mẫu

- Dữ liệu kiểm tra 3: 7156 mẫu

- Dữ liệu kiểm tra 4: 16695 mẫu

- Kết quả thực nghiệm

Accuracy = 0.99434855051

- Biểu đồ F-measure

- Nhận xét kết quả thực nghiệm và biểu đồ

+ Sau khi test trên nhiều bộ test khác nhau có thể thấy độ chênh lệch không nhiều độ chính xác cao nên thuật toán trên có thể dung được cho bộ dữ liệu này.