

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

---



**NGUYỄN THỊ HÒA**

**NGHIÊN CỨU MỘT SỐ THUẬT TOÁN HỌC MÁY VÀ  
ỨNG DỤNG**

**Chuyên ngành: Khoa học máy tính**

**Mã số: 60.48.01.01**

**TÓM TẮT LUẬN VĂN THẠC SĨ**

**HÀ NỘI - 2016**

Luận văn được hoàn thành tại:

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

Người hướng dẫn khoa học: TS Vũ Văn Thỏa.....  
(Ghi rõ học hàm, học vị)

Phản biện 1: TS. Phạm Văn Cường.....

Phản biện 2: PGS TS. Nguyễn Hải Châu .....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 9h giờ 00..... ngày 20..... tháng 08.... năm 2016.....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

## MỞ ĐẦU

Trong những năm gần đây, sự phát triển vượt bậc của công nghệ thông tin đã làm số lượng giao dịch thông tin trên mạng Internet tăng một cách đáng kể đặc biệt là thư viện điện tử, tin tức điện tử... Do đó mà số lượng văn bản xuất hiện trên mạng Internet cũng tăng với một tốc độ chóng mặt, và tốc độ thay đổi thông tin là cực kỳ nhanh chóng. Theo thống kê của Broder et al (2003) thì cứ sau 9 tháng hoặc 12 tháng lượng thông tin đó lại tăng gấp đôi. Cùng với đó là sự phổ cập máy tính và mạng internet, thói quen tìm kiếm thông tin qua mạng, đặc biệt là qua các trang web tìm kiếm nổi tiếng ngày càng phổ biến. Thông qua internet chúng ta có nhiều cơ hội để tiếp xúc với nguồn thông tin về vô cùng lớn. Nhưng cùng với nguồn thông tin vô tận đó, chúng ta cũng đang phải đối mặt với sự quá tải thông tin. Đồng thời độ tin cậy và chính xác của thông tin chưa cao. Đôi khi để tìm được thông tin cần thiết, chúng ta phải bỏ ra một lượng thời gian khá lớn, còn trong trường hợp chúng ta chưa rõ mình thực sự cần gì thì thời gian để tìm kiếm quả là không hề ít.

Với số lượng thông tin đồ sộ như vậy, một yêu cầu lớn đặt ra là làm sao tổ chức và tìm kiếm thông tin, dữ liệu có hiệu quả nhất. Giải pháp tác giả đưa ra là xây dựng các mô hình dự đoán dựa trên các phương pháp học máy và phân loại thông tin một cách tự động.

Xuất phát từ thực tế và mục tiêu như vậy, tác giả thực hiện đề tài luận văn có tên “Nghiên cứu một số thuật toán học máy và ứng dụng” để giải quyết vấn đề nêu trên.

### **Nội dung nghiên cứu:**

- Nghiên cứu một số kiến thức tổng quan về học máy.
- Nghiên cứu một số thuật toán học máy như cây quyết định, máy vectơ hỗ trợ

SVM, mạng nơ-ron nhân tạo

- Ứng dụng các thuật toán đã nghiên cứu để giải quyết bài toán phân loại cụ thể.

Qua đó, đánh giá độ chính xác và tính khả thi của thuật toán.

### **Nội dung luận văn gồm 3 chương:**

Chương 1: Tổng quan về học máy.

Chương 2: Nghiên cứu một số thuật toán học máy

Chương 3: Ứng dụng vào giải quyết bài toán phân loại

Trong đó đề tài tập trung vào chương 2 và 3 nhằm nghiên cứu tìm hiểu để đề xuất ứng dụng giải pháp phù hợp nhất với thực tế.

## Chương 1 - TỔNG QUAN VỀ HỌC MÁY

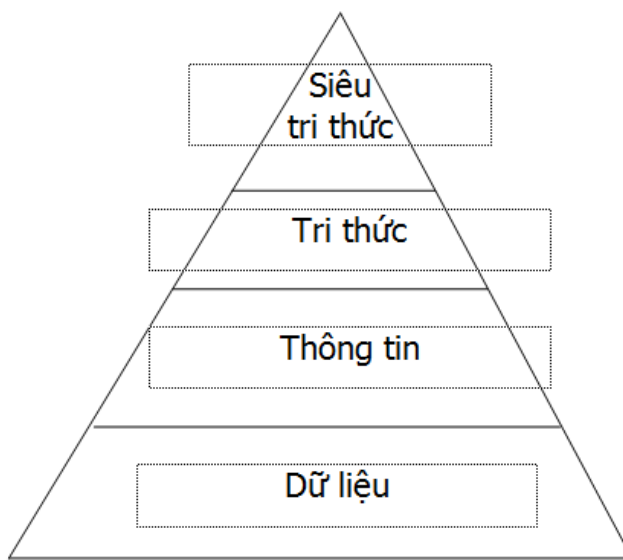
Chương này trình bày một số kiến thức tổng quan về học máy: những khái niệm cơ bản trong học máy, mô hình học máy, phân loại các phương pháp học máy, ứng dụng của học máy trong thực tế.

### 1.1 Một số khái niệm về học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể[36].

Rất khó để định nghĩa một cách chính xác về học máy. “Học - learn” có ý nghĩa khác nhau trong từng lĩnh vực: tâm lý học, giáo dục, trí tuệ nhân tạo,...

Một định nghĩa rộng nhất: “học máy là một cụm từ dùng để chỉ khả năng một chương trình máy tính để tăng tính thực thi dựa trên những kinh nghiệm đã trải qua” hoặc “học máy là để chỉ khả năng một chương trình có thể phát sinh ra một cấu trúc dữ liệu mới khác với các cấu trúc dữ liệu cũ”.



Hình 1.0.1 Mô hình kim tự tháp: Từ dữ liệu đến tri thức[32]

### 1.2 Phân loại các thuật toán học máy

Các thuật toán học máy được chia làm 3 loại: học có giám sát, học không giám sát và học nửa giám sát[32].

### **1.2.1 Học có giám sát**

Đây là cách học từ những mẫu dữ liệu mà ở đó các kỹ thuật học máy giúp hệ thống xây dựng cách xác định những lớp dữ liệu. Hệ thống phải tìm một sự mô tả cho từng lớp (đặc tính của mẫu dữ liệu). Người ta có thể sử dụng các luật phân loại hình thành trong quá trình học và phân lớp để có thể sử dụng dự báo các lớp dữ liệu sau này.

### **1.2.2 Học không giám sát**

Đây là việc học từ quan sát và khám phá. Hệ thống khai thác dữ liệu được ứng dụng với những đối tượng nhưng không có lớp được định nghĩa trước, mà để nó phải tự hệ thống quan sát những mẫu và nhận ra mẫu. Hệ thống này dẫn đến một tập lớp, mỗi lớp có một tập mẫu được khám phá trong tập dữ liệu. Học không giám sát còn gọi là học từ quan sát và khám phá.

### **1.2.3 Học bán giám sát**

Học bán giám sát là các thuật toán học tích hợp từ học giám sát và học không giám sát. Học bán giám sát sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện - điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn.

## **1.3 Ứng dụng của học máy**

Học máy có ứng dụng rộng khắp trong các ngành khoa học và sản xuất, đặc biệt những ngành cần phân tích khối lượng dữ liệu khổng lồ. Dưới đây là một số ứng dụng phổ biến của học máy.

### **1.3.1 Ứng dụng trong phân tích dự báo**

### **1.3.2 Ứng dụng trong tìm kiếm**

### **1.3.3 Ứng dụng trong phân lớp**

## **1.4 Kết chương**

Trong chương 1 luận văn đã khảo sát các vấn đề chung nhất của học máy. Các mô hình học máy được ứng dụng trong nhiều lĩnh vực khác nhau của đời sống xã hội như lĩnh vực y tế, sinh học, lĩnh vực kinh tế tài chính, quản trị kinh doanh,... Tuy nhiên, do đòi hỏi ngày càng cao của thực tiễn, học máy còn gặp nhiều thách thức và đang là một trong những lĩnh vực thu hút sự quan tâm của các nhà khoa học và các tổ chức cũng như doanh nghiệp.

Trong chương này luận văn cũng trình bày tổng quan về các học máy và một số ứng dụng của học máy trong các lĩnh vực. Có nhiều mô hình học máy, trong đó phương pháp

phân lớp được ứng dụng rất rộng rãi trong thực tế. Trong phương pháp phân lớp, kỹ thuật học máy SVM, cây quyết định, mạng nơ ron là những thuật toán phân loại được ứng dụng rộng rãi, đặc biệt là trong y học và tin sinh học.

Vì vậy, chương tiếp theo luận văn sẽ nghiên cứu ba thuật toán cơ bản là cây quyết định, SVM và mạng nơ ron.

## Chương 2: NGHIÊN CỨU MỘT SỐ THUẬT TOÁN HỌC MÁY

*Chương này trình bày một số thuật toán học máy tiêu biểu, cụ thể là thuật toán cây quyết định, vector hỗ trợ SVM và mạng nơron nhân tạo.*

### 2.1 Cây quyết định

Cây quyết định là một trong phương pháp học máy tiêu biểu có nhiều ứng dụng trong phân loại và dự đoán. Mặc dù độ chính xác của phương pháp này không thật cao so với những phương pháp được nghiên cứu gần đây, học cây quyết định vẫn có nhiều ưu điểm như đơn giản, dễ lập trình, và cho phép biểu diễn hàm phân loại dưới dạng dễ hiểu, dễ giải thích cho con người.

#### 2.1.1 Tổng quan về cây quyết định

##### 2.1.1.1 Định nghĩa

Cây quyết định là một cấu trúc ra quyết định có dạng cây. Cây quyết định nhận đầu vào là một bộ giá trị thuộc tính mô tả một đối tượng hay một tình huống và trả về một giá trị rời rạc. Mỗi bộ thuộc tính đầu vào được gọi là một mẫu hay một ví dụ, đầu ra gọi là loại hay nhãn phân loại. Thuộc tính đầu vào còn được gọi là đặc trưng và có thể nhận giá trị rời rạc hoặc liên tục. Để cho đơn giản, trước tiên ta sẽ xem xét thuộc tính rời rạc, sau đó sẽ mở rộng cho trường hợp thuộc tính nhận giá trị liên tục. Trong các trình bày tiếp theo, tập thuộc tính đầu vào được cho dưới dạng véc tơ  $x$ , nhãn phân loại đầu ra được ký hiệu là  $y$ , cây quyết định là hàm  $f(x)$  trả lại giá trị  $y$ . Cây quyết định được biểu diễn dưới dạng một cấu trúc cây (hình 2.1).

##### 2.1.1.2 Chiến lược cơ bản xây dựng cây quyết định

##### 2.1.1.3 Thuận lợi và hạn chế của mô hình cây quyết định

### 2.1.2 Thuật toán ID3

Giải thuật quy nạp cây quyết định ID3 (gọi tắt là ID3) là một giải thuật học đơn giản nhưng tỏ ra thành công trong nhiều lĩnh vực. ID3 là một giải thuật hay vì cách biểu diễn tri thức học được của nó, tiếp cận của nó trong việc quản lý tính phức tạp, heuristic của nó dùng cho việc chọn lựa các khái niệm ứng viên, và tiềm năng của nó đối với việc xử lý dữ liệu nhiễu.

ID3 biểu diễn các khái niệm ở dạng các cây quyết định. Biểu diễn này cho phép chúng ta xác định phân loại của một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó.

### 2.1.2.1 Thuật toán

Hàm xây dựng cây quyết định như sau:

```

Function induce_tree(tập_ví_dụ, tập_thuộc_tính)

  begin

    if mọi ví dụ trong tập_ví_dụ đều nằm trong cùng một lớp then

      return một nút lá được gán nhãn bởi lớp đó;

    else if tập_thuộc_tính là rỗng then return nút lá được gán nhãn bởi tuyến của
    tất cả các lớp trong tập_ví_dụ

    else begin

      chọn một thuộc tính P, lấy nó làm gốc cho cây hiện tại;
      xóa P ra khỏi tập_thuộc_tính;
      với mỗi giá trị V của P;

        begin

          tạo một nhánh của cây gán nhãn V;

          Đặt vào phân_vùng V các ví dụ trong tập_ví_dụ có giá trị V tại
          thuộc tính P;

          Gọi induce_tree(phân_vùng V , tập_thuộc_tính), gán kết quả
          vào nhánh V ;

        end;

      end ;

    end;
  
```



### 2.1.2.2 Thuộc tính phân loại tốt nhất

- a. Entropy đo tính thuần nhất của tập huấn luyện
- b. Lượng thông tin thu được đo mức độ giảm entropy mong đợi

### 2.1.2.3 Tìm kiếm không gian giả thuyết trong ID3

### 2.1.2.4 Đánh giá hiệu suất của cây quyết định:

### 2.1.2.5 Chuyển cây về các luật

## 2.1.3 Thuật toán C4.5

C4.5 là sự mở rộng của giải thuật ID3 trên một số khía cạnh sau:

- Trong việc xây dựng cây quyết định, chúng có thể liên hệ với tập huấn luyện mà có những bản ghi với những giá trị thuộc tính không được biết đến bởi việc đánh giá việc thu thập thông tin hoặc là tỉ số thu thập thông tin, cho những thuộc tính bằng việc xem xét chỉ những bản ghi mà ở đó thuộc tính được định nghĩa.
- Trong việc xây dựng cây quyết định, giải thuật C4.5 có thể giải quyết tốt đối với trường hợp giá trị của các thuộc tính là giá trị thực.
- Trong việc xây dựng cây quyết định, C4.5 có thể giải quyết tốt đối với trường hợp thuộc tính có nhiều giá trị mà mỗi giá trị này lại duy nhất.
- Trong việc sử dụng cây quyết định, chúng ta có thể phân loại những bản ghi mà có những giá trị thuộc tính không biết bằng việc ước lượng những kết quả có khả năng xảy ra.

## 2.2 Thuật toán máy véc tơ hỗ trợ SVM

### 2.2.1 Giới thiệu

SVM sử dụng thuật toán học nhằm xây dựng một siêu phẳng làm cực tiểu hoá độ phân lớp sai của một đối tượng dữ liệu mới. Độ phân lớp sai của một siêu phẳng được đặc trưng bởi khoảng cách bé nhất tới siêu phẳng đấy. SVM có khả năng rất lớn cho các ứng dụng được thành công trong bài toán phân lớp văn bản.

### 2.2.2 Định Nghĩa

### 2.2.3 Phương pháp SVM phân lớp nhị phân

Xét bài toán phân lớp nhị phân với tập dữ liệu mẫu huấn luyện

$$T = \{(x_i, y_i), i = 1, 2, \dots, n, x_i \in \mathbb{R}^d\},$$

Trong đó, các dữ liệu mẫu xi được biểu diễn dưới dạng véc tơ trong không gian véc tơ  $R_d$ . Các mẫu dương là các mẫu xi thuộc lĩnh vực quan tâm được gán nhãn  $y_i = +1$ ; các mẫu âm là các mẫu xi không thuộc lĩnh vực quan tâm được gán nhãn  $y_i = -1$ .

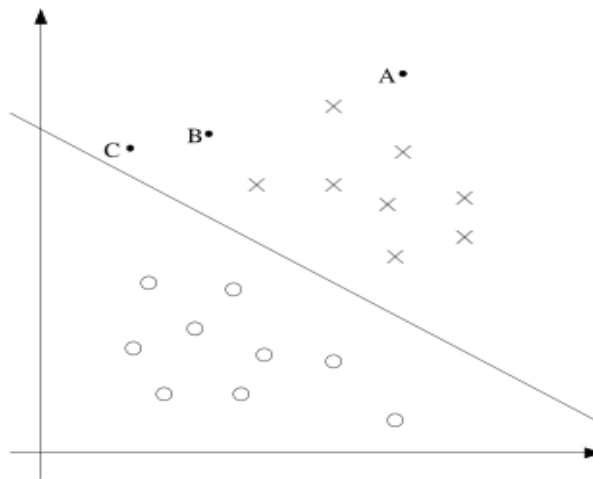
Khi đó cần tìm ra một ranh giới để phân tách các mẫu thành hai lớp tương ứng  $+1$  và  $-1$ . Độ chính xác của bộ phân lớp phụ thuộc vào độ lớn khoảng cách của điểm dữ liệu gần nhất của mỗi lớp đến ranh giới phân tách (còn gọi là ranh giới quyết định), khoảng cách đó còn gọi là biên.

Tùy thuộc vào dạng của ranh giới phân tách ta sẽ có SVM tuyến tính và SVM phi tuyến.

### 2.2.3.1 SVM tuyến tính

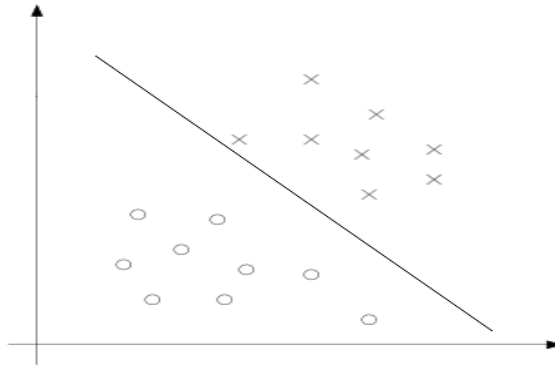
Trong không gian véc tơ  $R_d$  ta sẽ xác định ranh giới phân tách hai lớp có dạng là một siêu phẳng.

Để rõ hơn về tầm quan trọng của biên đối với siêu phẳng phân tách (siêu phẳng quyết định) ta xét ví dụ sau đây (Hình 2.6).

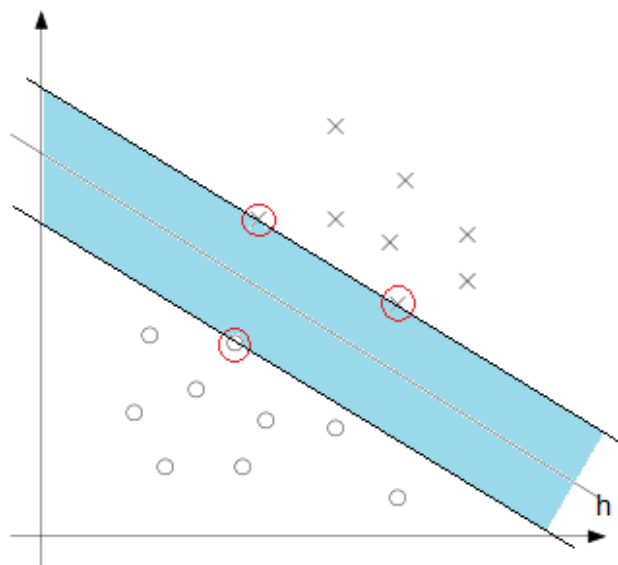


**Hình 2.1 Tầm quan trọng của biên trong thuật toán SVM**

Trong hình 2.4, ta có thể nhận thấy rằng, các điểm có khoảng cách tới siêu phẳng phân tách lớn như điểm A thì có thể gán A vào lớp  $+1$  mà không sợ có sai sót. Trong khi đó, với điểm C ngay sát siêu phẳng phân tách sẽ được dự đoán thuộc lớp  $+1$  nhưng C cũng có thể thuộc lớp  $-1$  nếu có một sự thay đổi nhỏ của siêu phẳng phân tách. Điểm B nằm giữa hai trường hợp này. Như vậy, khoảng cách biên càng lớn thì siêu phẳng quyết định càng tốt và độ chính xác phân loại càng cao. Mục đích của SVM là tìm ra siêu phẳng có khoảng cách biên lớn nhất, còn gọi là siêu phẳng tối ưu.



**Hình 2.2 Ví dụ về một biên không tốt**



**Hình 2.3 Ví dụ về biên tối ưu**

### 2.2.3.2 SVM phi tuyến tính

Trong thực tế các tập dữ liệu huấn luyện có ranh giới quyết định là không tuyến tính vì vậy rất khó giải quyết. Tuy nhiên chúng ta có thể chuyển tập dữ liệu huấn luyện này về dạng tuyến tính quen thuộc bằng cách ánh xạ dữ liệu này sang một không gian có số chiều lớn hơn gọi là không gian đặc trưng (feature space). Với không gian đặc trưng phù hợp thì dữ liệu huấn luyện sau khi ánh xạ sẽ trở nên khả tuyến và phân tách dữ liệu sẽ ít lỗi hơn so với không gian ban đầu. Phương pháp SVM phi tuyến có thể phân thành hai bước như sau:

Bước 1: Chuyển đổi không gian dữ liệu ban đầu sang một không gian đặc trưng khác (thường có số chiều lớn hơn), khi đó dữ liệu huấn luyện có thể phân tách tuyến tính được.

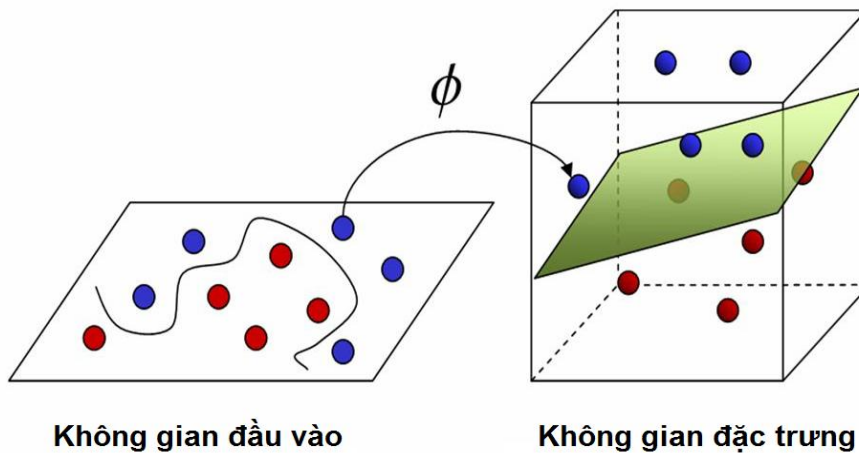
Bước 2: Áp dụng các công thức như với SVM tuyến tính.

Giả sử dữ liệu  $x_i$  ban đầu thuộc không gian  $R^d$  ta sử dụng một hàm ánh xạ  $\phi$  để chuyển tập dữ liệu  $x_i$  sang không gian  $R^m$ .

$$\phi: R^d \rightarrow R^m, x \mapsto \phi(x)$$

Tập huấn luyện  $T$  ban đầu được ánh xạ thành tập

$$T' = \{(\phi(x_1), y_1), (\phi(x_2), y_2), \dots, (\phi(x_n), y_n))\}$$



**Hình 2.4 Ánh xạ từ không gian 2 chiều sang không gian 3 chiều**

### 2.2.3.3 Thuật toán tối thiểu tuần tự SMO

Cả hai bài toán gốc và bài toán đối ngẫu của thuật toán SVM đều là bài toán tối ưu bậc 2 (Quadratic Programming) và đều có thể giải bằng phương pháp điểm trong (interior-point methods). Tuy nhiên khi số lượng mẫu học  $n$  lớn thì ma trận  $K$  cũng lớn lên theo bậc 2 của  $n$ . Vì vậy phương pháp điểm trong cũng có thời gian chạy rất lâu cỡ  $O(n^3)$ . Vì vậy, ta phải lợi dụng cấu trúc của bài toán tối ưu trong thuật toán SVM để tăng tốc độ tối ưu hóa.

#### **Thuật toán tối thiểu tuần tự (Sequential Minimal Optimization – SMO)**

Đây là thuật toán tối ưu dành riêng cho phương pháp SVM do J. Platt đưa ra vào năm 1998. Ý tưởng chính của thuật toán này là:

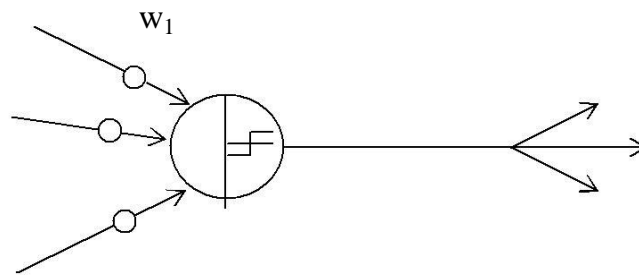
- Thay vì không chế tất cả các ràng buộc, ta cố định phần lớn các biến  $\lambda_i$  và chỉ tối ưu hóa một cặp  $(\lambda_i, \lambda_j)$  nào đó.
- Giá trị tối ưu của cặp  $(\lambda_i, \lambda_j)$  có thể viết dưới dạng công thức (của dữ liệu và các biến  $\lambda_i$  khác) chứ không cần chạy một thuật toán tối ưu nào cả.
- Lần lượt chọn các cặp  $(\lambda_i, \lambda_j)$  theo một tiêu chí (heuristics) nào đó để thuật toán nhanh chóng hội tụ về nghiệm tối ưu.

Thuật toán tối thiểu tuần tự SMO được sử dụng trong hầu hết tất cả bài toán cài đặt thuật toán SVM.

## 2.3 Thuật toán mạng nơ ron nhân tạo

### 2.3.1 Giới thiệu

Mạng nơ ron nhân tạo là một mô phỏng xử lý thông tin, được nghiên cứu ra từ hệ thống thần kinh của sinh vật, giống như bộ não để xử lý thông tin. Nó bao gồm số lượng lớn các môi gắn kết cấp cao để xử lý các yếu tố làm việc trong môi liên hệ giải quyết vấn đề rõ ràng. ANNs giống như con người, được học bởi kinh nghiệm, lưu những kinh nghiệm hiểu biết và sử dụng trong những tình huống phù hợp.

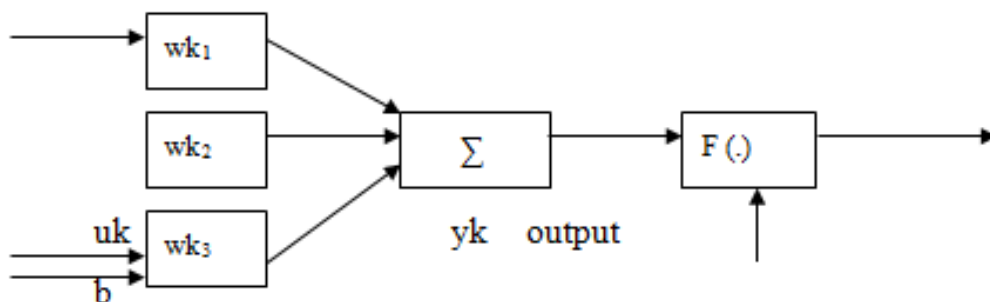


Hình 2.5 Mô hình mạng nơ ron nhân tạo

### 2.3.2 Cơ sở lý thuyết

#### 2.3.2.1 Cấu trúc mạng nơ ron

Mỗi Neural (nút) là một đơn vị xử lý thông tin của mạng neural, là yếu tố cơ bản để cấu tạo nên mạng neural.



Hình 2.6 Cấu trúc 1 nơ ron (Neural)

$x_i$ : các tín hiệu input

wkp: trọng số của từng input

$f(\cdot)$ : hàm hoạt động

yk: kết xuất của Neural

b: thông số ảnh hưởng đến ngưỡng ra của output

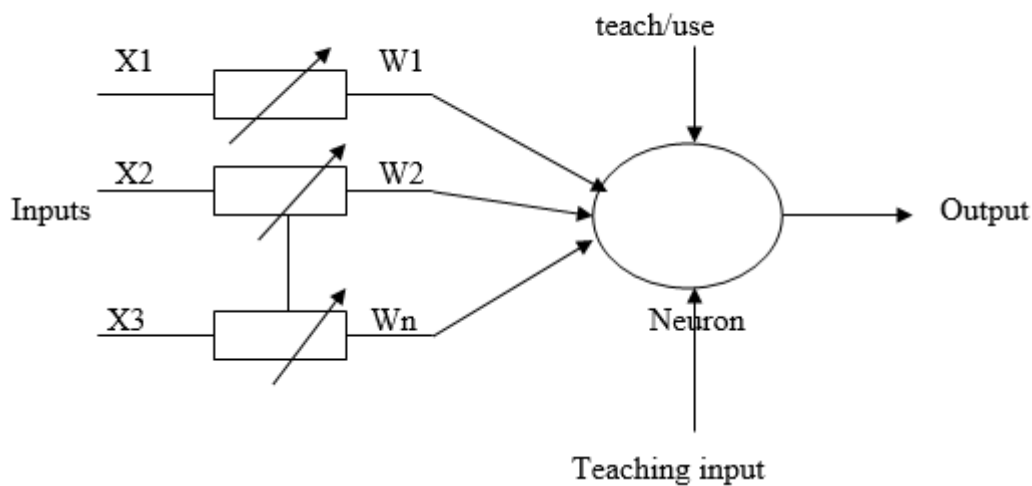
a. Mạng dẫn tiến một lớp

b. Mạng dẫn tiến nhiều lớp

### 2.3.2.2 Khả năng ứng dụng của mạng nơ-ron nhân tạo

### 2.3.2.3 Tiến trình học

Tiến trình học là tiến trình quan trọng của con người, nhờ học mà bộ não ngày càng tích lũy những kinh nghiệm để thích nghi với môi trường và xử lý tình huống tốt hơn. Mạng neural xây dựng lại cấu trúc bộ não thì cần phải có khả năng nhận biết dữ liệu thông qua tiến trình học, với các thông số tự do của mạng có thể thay đổi liên tục bởi những thay đổi của môi trường và mạng neural ghi nhớ giá trị đó.



**Hình 2.7 Tiến trình học**

Trong quá trình học, giá trị đầu vào được đưa vào mạng và theo dòng chảy trong mạng tạo thành giá trị ở đầu ra. Tiếp đến là quá trình so sánh giá trị tạo ra bởi mạng nơ ron với giá trị ra mong muốn. Nếu hai giá trị này giống nhau thì không thay đổi gì cả. Tuy nhiên, nếu có một sai lệch giữa hai giá trị này vượt quá giá trị sai số mong muốn thì đi ngược mạng từ đầu ra về đầu vào để thay đổi một số kết nối.

### 2.3.2.4 Giải thuật Back – Propagation

Thuật toán Back – Propagation được sử dụng để điều chỉnh các trọng số kết nối sao cho tổng sai số  $e$  nhỏ nhất.

$$E = \sum_{i=1}^n (t(x_i, w) - y(x_i))^2$$

Trong đó:

$t(x_i, w)$ : giá trị của tập mẫu

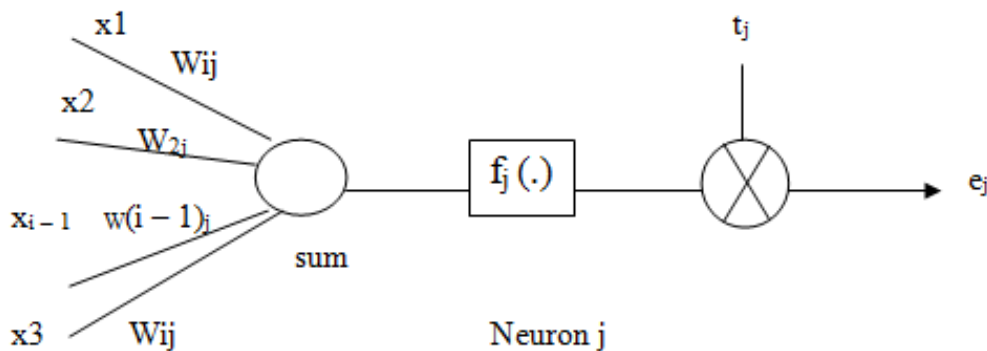
$y(x_i)$ : giá trị kết xuất của mạng

Trước tiên, ta xét trên 1 Neuron, mỗi Neuron đều có giá trị vào và ra, mỗi giá trị đều có một trọng số để đánh giá mức độ ảnh hưởng của giá trị vào đó. Thuật toán Back – Propagation sẽ điều chỉnh các trọng số đó để giá trị  $e_j = T_j - y_j$  là nhỏ nhất.

Trước hết ta phải xác định vị trí của mỗi neuron. Neuron nào là của lớp ẩn và neuron nào là của lớp xuất. Ta cần biết các ký hiệu:

$W_{ij}$ : vector trọng số của neuron  $j$  số đầu vào  $i$

$u_j$ : vector giá trị kết xuất của neuron trong lớp  $j$



Hình 2.8 Mô hình tính toán một neuron

### 2.3.2.5 Những hạn chế của phương pháp lan truyền ngược

## 2.4 So sánh các thuật toán

	Máy véc tơ hỗ trợ SVM	Cây quyết định	Mạng nơ-ron
Ưu điểm	- Tối ưu toàn cục, mô hình chất lượng cao, chịu đựng	- Cây quyết định có thể xử lý cả thuộc tính tên và số đầu vào.	- Có khả năng mô phỏng các hàm cực kỳ phức tạp.

	<p>được nhiều.</p> <ul style="list-style-type: none"> <li>- SVM là một phương pháp tốt (phù hợp) đối với những bài toán phân loại có không gian biểu diễn thuộc tính lớn. Các đối tượng cần phân loại được biểu diễn bởi một tập rất lớn các thuộc tính.</li> </ul>	<ul style="list-style-type: none"> <li>- Thể hiện của cây quyết định là đủ đa dạng để biểu diễn cho bất kỳ giá trị rời rạc nào.</li> <li>- Cây quyết định có khả năng xử lý các bộ dữ liệu mà có thể gây ra lỗi.</li> <li>- Cây quyết định có khả năng xử lý các bộ dữ liệu mà có giá trị rỗng.</li> </ul>	<ul style="list-style-type: none"> <li>- Mạng nơ-ron nhân tạo có thể học từ những dữ liệu huấn luyện và khái quát những tình huống mới.</li> <li>- Có khả năng chịu lỗi, nhiều dữ liệu.</li> </ul>
Nhược điểm	<ul style="list-style-type: none"> <li>- SVM chỉ làm việc với không gian đầu vào là các số thực. Đối với các thuộc tính định danh (nominal), cần chuyển các giá trị định danh thành các giá trị số</li> <li>- Độ phức tạp vẫn cao</li> <li>- Xử lý dữ liệu kiểu số</li> </ul>	<ul style="list-style-type: none"> <li>- Hầu hết các thuật toán bắt buộc các thuộc tính mục tiêu phải là các giá trị rời rạc.</li> <li>- SVM chỉ làm việc (thực hiện phân loại) với 2 lớp. Đối với các bài toán phân loại nhiều lớp, cần chuyển thành m thành một tập các bài toán phân loại gồm 2 lớp, và sau đó giải quyết riêng rẽ từng bài toán 2 lớp này</li> </ul>	<ul style="list-style-type: none"> <li>- Chỉ xử lý được dữ liệu số trong những khoảng thích hợp cho mạng,</li> <li>- Khó xử lý với dữ liệu định danh.</li> <li>- Dữ liệu huấn luyện ít sẽ dẫn đến không đủ thông tin để huấn luyện mạng.</li> </ul>

## 2.5 Kết chương

Chương 2 trình bày cơ sở lý thuyết về một số thuật toán học máy cơ bản là thuật toán cây quyết định, thuật toán SVM, giải thuật mạng nơ-ron nhân tạo. Với khả năng vượt trội



của mỗi thuật toán về tính hiệu quả, độ chính xác, khả năng xử lý các bộ dữ liệu một cách linh hoạt, việc sử dụng các thuật toán học máy đã và đang là sự lựa chọn tối ưu nhất trong việc giải quyết các bài toán phân loại/dự báo trong một số các ngành khoa học, đặc biệt là y học lâm sàng.

Trên cơ sở lý thuyết của thuật toán SVM và cây quyết định, chương 3 xây dựng mô hình phân loại tổng quát và thử nghiệm trên tập dữ liệu bệnh dựa trên kết quả xét nghiệm hóa nghiệm và các triệu chứng ban đầu được bác sĩ chẩn đoán.

## Chương 3: ỨNG DỤNG GIẢI QUYẾT BÀI TOÁN PHÂN LỚP

*Chương này giới thiệu về bài toán phân loại và phương pháp phân loại dữ liệu. Trên cơ sở đó, luận văn sử dụng kỹ thuật SVM và cây quyết định để xây dựng mô hình phân loại cho bài toán cụ thể đó là phân loại bệnh dựa trên cơ sở phát hiện triệu chứng lâm sàng và kết quả xét nghiệm, hóa nghiệm.*

### 3.1 Bài toán phân lớp

#### 3.1.1 Giới thiệu

Bài toán phân loại là việc gán các nhãn phân loại cho dữ liệu mới dựa trên mức độ tương tự của tập dữ liệu đó so với các dữ liệu đã được gán nhãn trong tập huấn luyện. Nhiều kỹ thuật máy học và khai phá dữ liệu đã được áp dụng vào bài toán phân loại, chẳng hạn: phương pháp Naive Bayes, cây quyết định, k–láng giềng gần nhất (KNN), mạng nơron (neural network),... Máy học vector hỗ trợ (SVM) là một giải thuật phân lớp có hiệu quả cao và đã được áp dụng nhiều trong lĩnh vực khai phá dữ liệu và nhận dạng. Trong luận văn này nghiên cứu thuật toán máy vector hỗ trợ (SVM), áp dụng nó vào bài toán phân lớp và so sánh hiệu quả của nó với hiệu quả của giải thuật phân lớp cổ điển, rất phổ biến đó là cây quyết định. Nghiên cứu chỉ ra rằng SVM với cách lựa chọn đặc trưng bằng phương pháp tách giá trị đơn (SVD) cho kết quả tốt hơn so với cây quyết định.

#### 3.1.2 Mô tả bài toán phân lớp

Cho tập các mẫu đã phân lớp trước, xây dựng mô hình cho từng lớp. Mục đích là gán các mẫu mới vào các lớp với độ chính xác cao nhất có thể.

Cho cơ sở dữ liệu  $D = \{t_1, t_2, \dots, t_n\}$  và tập các lớp  $C = \{C_1, \dots, C_m\}$ , phân lớp là bài toán xác định ánh xạ  $f: D \rightarrow C$  sao cho mỗi  $t_i$  được gán vào một lớp.

#### 3.1.3 Phương pháp phân lớp

Quy trình phân lớp:

Bước 1: xây dựng mô hình phân lớp

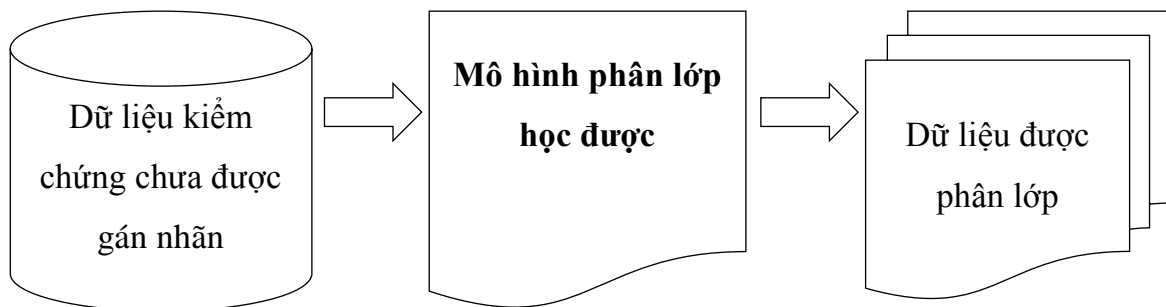
- Mô tả tập các lớp xác định trước bao gồm tập huấn luyện (các mẫu/ các bộ) dành cho xây dựng mô hình. Mỗi mẫu/ bộ thuộc một lớp đã được định trước.
- Tìm luật phân lớp, cây quyết định hoặc công thức toán mô tả lớp.



**Hình 3.1 Giai đoạn xây dựng mô hình**

**Bước 2:** Sử dụng mô hình

- Phân lớp các đối tượng chưa biết
  - Xác định độ chính xác của mô hình. Tập dữ liệu kiểm tra độc lập với tập dữ liệu huấn luyện gọi là tập kiểm chứng để kiểm định mô hình.
  - Độ chính xác chấp nhận được. Áp dụng mô hình để phân lớp các mẫu chưa xác định được nhãn.

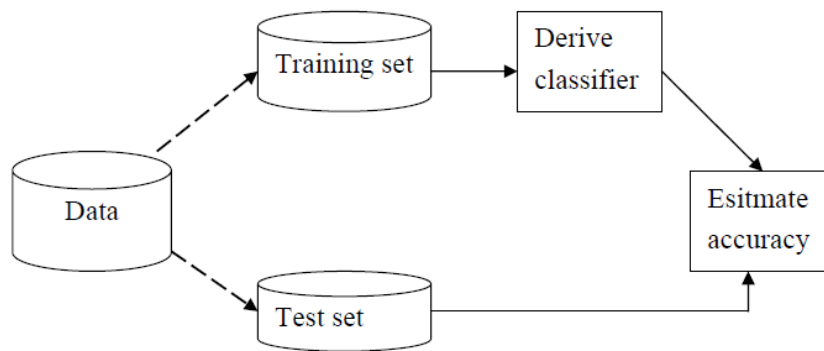


**Hình 3.2 Giai đoạn phân lớp**

Độ chính xác của mô hình trên tập kiểm chứng đã đưa là tỉ lệ phần trăm các các bộ trong tập dữ liệu kiểm tra được mô hình phân lớp đúng (so với thực tế). Nếu độ chính xác của mô hình là chấp nhận được, thì mô hình được sử dụng để phân lớp những dữ liệu tương lai, hoặc những dữ liệu mà giá trị của thuộc tính phân lớp là chưa biết.

### **3.1.4 Đánh giá mô hình**

Ước lượng độ chính xác của bộ phân lớp là quan trọng ở chỗ nó cho phép dự đoán được độ chính xác của các kết quả phân lớp những dữ liệu tương lai. Độ chính xác còn giúp so sánh các mô hình phân lớp khác nhau..

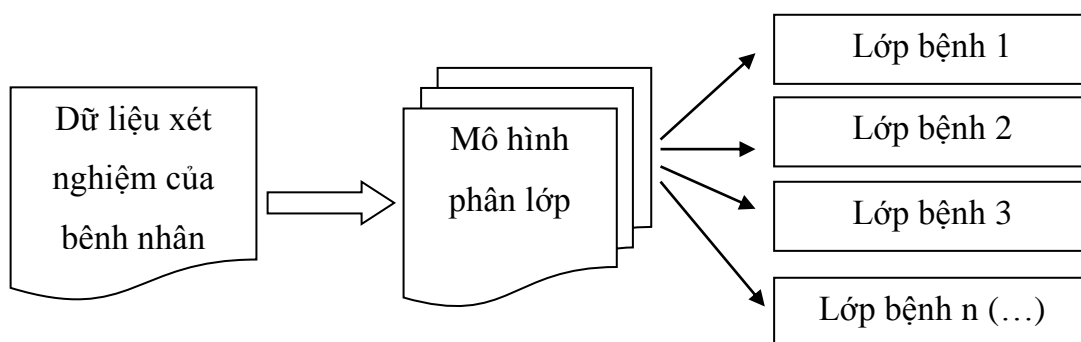


**Hình 3.3 Đánh giá độ chính xác của mô hình phân lớp**

## 3.2 Bài toán phân loại bệnh dựa trên dấu hiệu khám bệnh lâm sàng và các chỉ số xét nghiệm hóa nghiệm

### 3.2.1 Đặt bài toán

Bài toán đặt ra là: Cho trước một mẫu dữ liệu về một số bệnh phổ biến và các triệu chứng lâm sàng của bệnh nhân sử dụng phương pháp SVM và cây quyết định xây dựng mô hình phân lớp để xác định mẫu đó thuộc lớp bệnh đã có nào và so sánh hiệu quả của hai mô hình phân loại (Hình 3.4).



**Hình 3.4 Mô hình bài toán phân lớp mặt bệnh**

### 3.2.2 Các bước giải bài toán

Phương pháp giải bài toán theo mô hình trong hình 3.2 được thực hiện với các bước như sau (Hình 3.5):

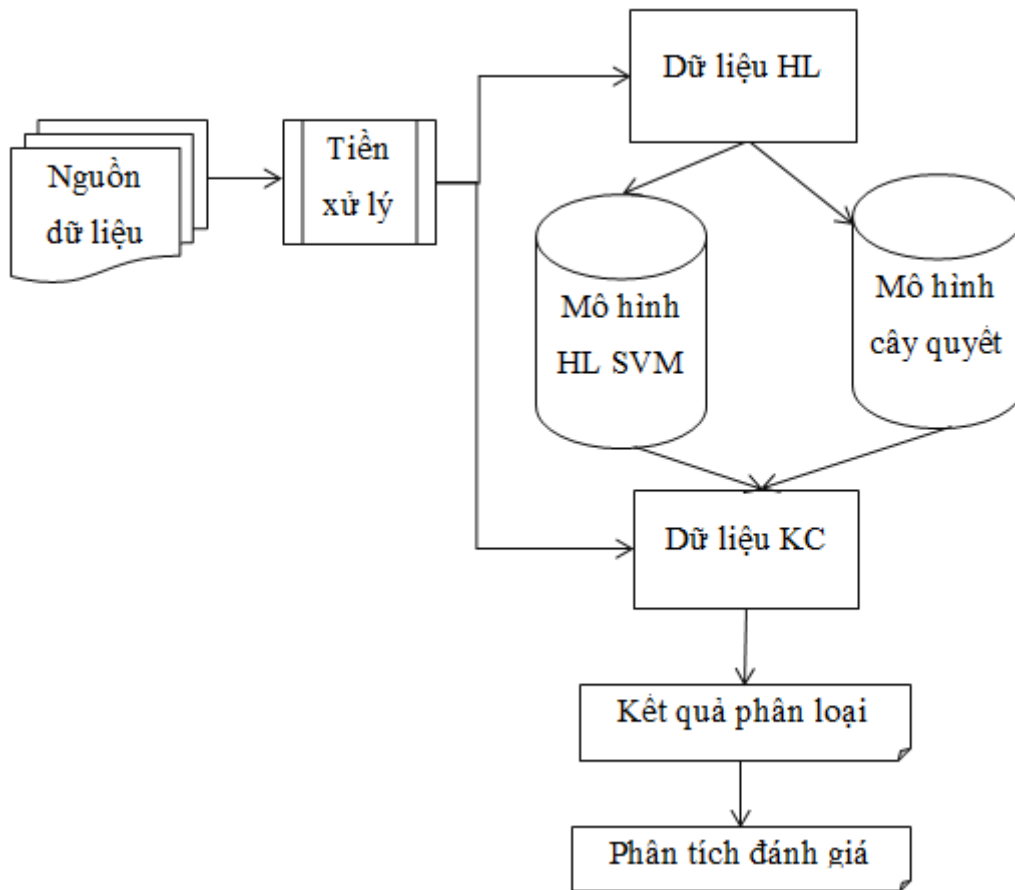
Bước 1: Thu thập dữ liệu xét nghiệm;

Bước 2: Tiền xử lý dữ liệu;

Bước 3: Phân chia dữ liệu thành tập dữ liệu huấn luyện và tập dữ liệu kiểm chứng;

Bước 4: Xây dựng mô hình phân lớp theo SVM và cây quyết định trên tập huấn luyện;

Bước 5: Sử dụng mô hình phân lớp có được để phân loại với tập dữ liệu kiểm chứng và đánh giá kết quả của mô hình.



**Hình 3.5 Các bước phân lớp mắt bệnh dựa trên triệu chứng lâm sàng và cận lâm sàng**

### 3.2.3 Thu thập dữ liệu nghiên cứu

Tiêu chuẩn đặt ra để lựa chọn mắt bệnh là:

- Những mắt bệnh phổ biến được điều trị tại một số bệnh viện đa khoa tuyến tỉnh và tuyến huyện
- Những mắt bệnh không do tác động của lực và thuộc cơ quan cụ thể của người (ví dụ: Sốt, nhiễm khuẩn, nhiễm trùng là bệnh về toàn thân, chấn thương, vết thương là những bệnh do tác động của lực nên không được lựa chọn)
- Có thể xác định bệnh qua các triệu chứng lâm sàng và một vài xét nghiệm liên quan.

Qua thu thập thông tin và tổng hợp đã lựa chọn được 800 bệnh nhân phù hợp tiêu chuẩn lựa chọn với phân bố theo bảng 3.1 sau:

**Bảng 3.1 Số lượng BN theo nhóm mặt bệnh nghiên cứu**

Mã nhóm	Tên nhóm mặt bệnh	Số lượng (n)
1	Nhóm bệnh lý về đường hô hấp	248
2	Nhóm bệnh lý khớp	82
3	Nhóm bệnh tim mạch	236
4	Nhóm bệnh lý đái tháo đường	286
	TỔNG CỘNG	852

### 3.2.4 Tiền xử lý dữ liệu

Việc tiền xử lý dữ liệu cho quá trình phân lớp là một việc làm không thể thiếu và có vai trò quan trọng quyết định tới sự áp dụng được hay không của mô hình phân lớp. Quá trình tiền xử lý dữ liệu sẽ giúp cải thiện độ chính xác, tính hiệu quả và khả năng mở rộng được của mô hình phân lớp.

#### a. Làm sạch dữ liệu:

Làm sạch dữ liệu liên quan đến việc xử lý nhiễu (noise) và giá trị thiếu (missing value) trong tập dữ liệu ban đầu.

#### b. Lựa chọn, rút gọn thuộc tính

Lựa chọn thuộc tính (Feature Selection, Feature Extraction) là nhiệm vụ rất quan trọng giai đoạn tiền xử lý dữ liệu khi triển khai các mô hình khai phá dữ liệu. Một vấn đề gặp phải là các tập dữ liệu dùng để xây dựng các mô hình phân lớp thường chứa nhiều thông tin không cần thiết (thậm chí gây nhiễu) cho việc xây dựng mô hình làm giảm độ chính xác của mô hình và gây khó khăn trong việc phát hiện tri thức.

### 3.2.5 Bài toán thực nghiệm

Ký hiệu tập các đặc trưng (các chỉ số) xét nghiệm hóa nghiệm là  $F = \{f_1, f_2, \dots, f_d\}$  với  $d = 46$ . Kết quả xét nghiệm và chẩn đoán bệnh của mỗi bệnh nhân  $p_i$  sẽ được biểu diễn bằng một véc tơ trong không gian  $R_d$ :  $p_i = \{f_{i1}, f_{i2}, \dots, f_{i46}\}$ ,  $f_{ij} \rightarrow R$  là giá trị kết quả của xét nghiệm  $f_j$  của bệnh nhân  $p_i$ .

Gọi tập  $C = \{c1, c2, ..., c4\}$  chứa các mã lớp nhóm bệnh cần phân loại tương ứng 4 nhóm bệnh kể trên.

Bài toán

Đầu vào:

- Tập T chứa n mẫu xét nghiệm huấn luyện đã mô hình hóa thành các véc tơ  $pi(fi1, fi2, ..., fi46)$ ;

- Tập F =  $\{f1, f2, ..., f46\}$  chứa mã chỉ số xét nghiệm hoặc các triệu chứng lâm sàng;

- Tập C =  $\{c1, c2, ..., c4\}$  chứa các mã lớp nhóm bệnh cần phân loại;

Đầu ra: Bộ phân lớp sử dụng SVM; Bộ phân lớp sử dụng cây quyết định ID3

### **Thuật toán sử dụng**

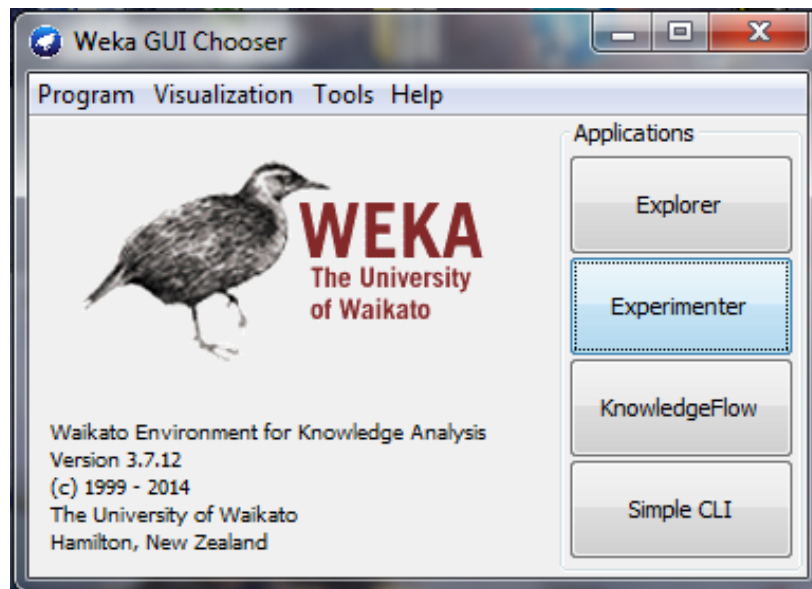
Để giải bài toán trên, học viên sử dụng SVM trong phân lớp đa lớp và cây quyết định để so sánh đánh giá kết quả.

## **3.3 Thử nghiệm và đánh giá kết quả**

### **3.3.1 Công cụ thực nghiệm**

Công cụ thực nghiệm: Sử dụng phần mềm Weka version 3.7.12.

Weka là một phần mềm miễn phí về học máy được viết bằng Java, phát triển bởi University of Wekato. Weka có thể coi như là bộ sưu tập các thuật toán về học máy dùng trong phân tích và khai phá dữ liệu. Các thuật toán đã được xây dựng sẵn chỉ việc sử dụng. Do đó Weka rất thích hợp cho việc thử nghiệm các mô hình mà không mất thời gian để xây dựng chúng. Weka có giao diện sử dụng đồ họa trực quan và cả chế độ command line. Ngoài các thuật toán về học máy như dự đoán, phân loại, phân cụm, Weka còn có các công cụ để trực quan hóa dữ liệu rất hữu ích trong quá trình nghiên cứu, phân tích.



**Hình 3.6** Giao diện khởi động của WEKA

### 3.3.2 Chuẩn bị dữ liệu

Mẫu XN gồm 852 mẫu thuộc 04 nhóm bệnh đã được tiền xử lý.

**Bảng 3.2** Cơ cấu số mẫu HL và KC tương ứng

TT	Tên nhóm mặt bệnh	Số lượng (n)	Số lượng HL	Số lượng KC
	<b>Nhóm bệnh</b>	<b>850</b>	<b>563</b>	<b>289</b>
1	Nhóm bệnh lý về đường hô hấp	248	177	71
2	Nhóm bệnh lý khớp	82	50	32
3	Nhóm bệnh lý tim mạch	236	142	94
4	Nhóm bệnh lý đái tháo đường	284	185	99

Dữ liệu được tách ra, chuyển đổi sang dạng chuẩn csv của Weka với 45 thuộc tính gồm 44 thuộc tính chỉ số XN và triệu chứng lâm sàng với 1 thuộc tính lớp theo bảng 3.4 như sau:

**Bảng 3.3** Cơ cấu của các tập tin dữ liệu

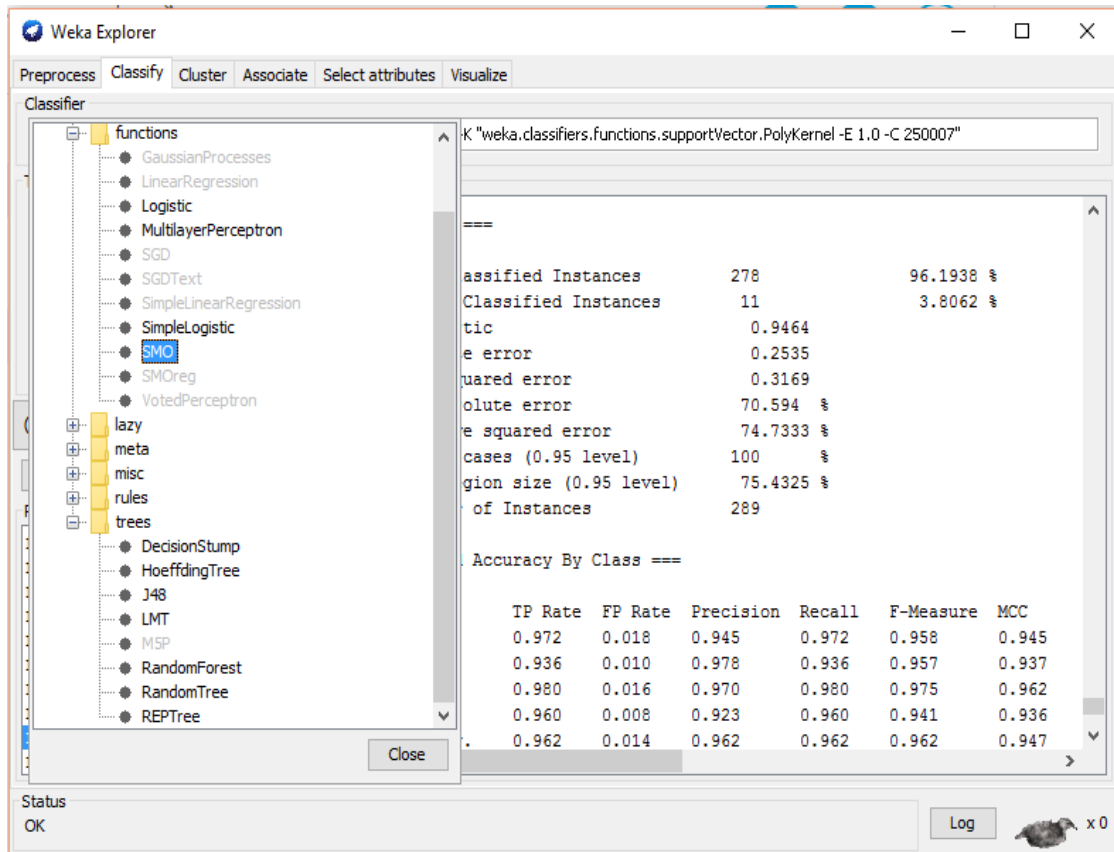
Tên tập tin	Nội dung	Số thuộc tính	Số bản ghi
DL_HL.csv	Tập huấn luyện	<b>46</b>	<b>563</b>



Tên tập tin	Nội dung	Số thuộc tính	Số bản ghi
DL_KC.csv	Tập kiểm chứng	46	289

### 3.3.3 Thực hiện thực nghiệm

Mỗi một thuật toán đều được thực hiện liên tiếp 5 lần đối với tập mẫu, mỗi lần thực hiện đều theo quy trình thực hiện từ bước 1.



**Hình 3.7 Thực hiện phân loại với J48 Classifier và SMO Classifier**

### 3.3.4 Kết quả thực nghiệm

Kết quả thử nghiệm:

**Bảng 3.4 Kết quả phân lớp theo cây quyết định J48**

```
=== Confusion Matrix ===
  a  b  c  d  <-- classified as
69  2  0  0 | a = HH
 5 86  3  0 | b = MN
 0  0 97  2 | c = O24.3
 0  0  4 21 | d = TK
```

**Bảng 3.5 Kết quả phân lớp theo thuật toán SMO**

```
=== Confusion Matrix ===
  a  b  c  d  <-- classified as
69  2  0  0 | a = HH
 4 88  2  0 | b = MN
 0  0 97  2 | c = O24.3
 0  0  1 24 | d = TK
```

**Bảng 3.6 Kết quả đánh giá thuật toán cây quyết định J48**

```
=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure
0.972    0.023    0.932    0.972    0.952
0.915    0.010    0.977    0.915    0.945
0.980    0.037    0.933    0.980    0.956
0.840    0.008    0.913    0.840    0.875
Weighted Avg.  0.945    0.022    0.945    0.945    0.944
```

**Bảng 3.7 Kết quả đánh giá thuật toán cây quyết định SMO**

```
=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure
0.972    0.018    0.945    0.972    0.958
0.936    0.010    0.978    0.936    0.957
0.980    0.016    0.970    0.980    0.975
0.960    0.008    0.923    0.960    0.941
Weighted Avg.  0.962    0.014    0.962    0.962    0.962
```

### 3.3.5 Phân tích và đánh giá kết quả

Qua các kết quả thực nghiệm nhận thấy:

Trên bảng 3.8 cho thấy lớp bệnh được phân loại với độ tin cậy cao nhất là lớp bệnh lý tim mạch với độ nhạy = 93.6,8 %, độ đặc hiệu = 99 % và độ chính xác = 96.3% ở kết quả phân loại sử dụng SMO. Có 2 lớp bệnh lý đái tháo đường và bệnh về đường hô hấp có độ tin cậy phân loại thấp hơn, theo đó độ nhạy, độ đặc hiệu và độ chính xác theo thuật toán SMO và J48 ở cả 2 lớp bệnh này lần lượt là: 98 % - 98.4% - 98.2% / 98% - 96.3.0% - 97.15% ở bệnh đái tháo đường và 97.2 % - 98.2% - 97.7%/97.2% - 97.7%-97.45% ở bệnh về đường hô hấp .

### **3.4 Kết luận chương**

Trong chương này luận văn đã đưa ra bài toán phân loại tổng quát và nêu lên phương pháp xây dựng mô hình phân loại. Luận văn đã khảo sát các khía cạnh của bài toán phân loại bệnh dựa trên triệu chứng lâm sàng ban đầu của bệnh nhân kết hợp với kết quả xét nghiệm hóa nghiệm. Trên cơ sở các dữ liệu thu thập được của 4 loại bệnh của các bệnh nhân tại một số bệnh viện hạng 3 và hạng 2, luận văn đã tiến hành thực nghiệm với việc sử dụng SVM và thuật toán cây quyết định. Kết quả thực nghiệm thu được khi sử dụng phần mềm WEKA được phân tích và đánh giá cho thấy sự phù hợp với lý thuyết đã nghiên cứu.

## KẾT LUẬN

### 1. Những đóng góp của luận văn:

Qua nghiên cứu và thực nghiệm, luận văn đã đạt được những kết quả chính như sau:

- Nghiên cứu tổng quan về học máy, các khái niệm cơ bản trong học máy và ứng dụng
- Nghiên cứu một số thuật toán học máy tiêu biểu đó là thuật toán SVM, thuật toán cây quyết định và mạng nơ-ron nhân tạo
- Ứng dụng các thuật toán đã tìm hiểu để giải quyết bài toán phân lớp thông qua các mô hình huấn luyện của các thuật toán đã tìm hiểu trên
- Đã thu thập và chuẩn hóa được bộ số liệu của 46 chỉ số xét nghiệm và triệu chứng

lâm sàng với trên 800 mẫu của 04 nhóm bệnh.

Đã tiến hành thực nghiệm và phân tích, đánh giá kết quả thu được. Bước đầu cho thấy ứng dụng SVM trong việc phân loại bệnh dựa trên các triệu chứng lâm sàng ban đầu và kết quả xét nghiệm hóa nghiệm đạt hiệu quả khả quan.

### 2. Hướng phát triển luận văn:

Tuy đạt được một số kết quả nêu trên, nhưng luận văn còn một số hạn chế do điều kiện về mặt thời gian và trình độ của học viên. Vì vậy, hướng nghiên cứu tiếp theo của học viên là:

- Nghiên cứu thêm về các thuật toán học máy khác, nghiên cứu thuật toán SVM, cây quyết định kết hợp với các thuật toán khác để có thể tăng độ chính xác phân lớp..
- Để ứng dụng trong thực tế có thể phát triển bài toán cụ thể trong chương ba bằng cách phân loại đối với bệnh nhân đa bệnh kết hợp (phân lớp đa nhãn).
- Mở rộng thêm mặt bệnh phân loại và phân loại mặt bệnh chi tiết hơn.
- Phát triển bài toán phân loại bệnh thành ứng dụng để có thể hỗ trợ định hướng chẩn đoán bệnh trong bệnh viện.