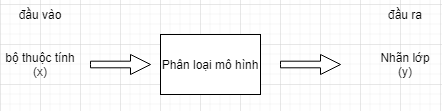
1. Dịch:

-Nếu cho các tập dữ liệu nhỏ và đơn giản thì việc phân loại thủ công là đủ với một vài thuộc tính, nhưng với tập dữ liệu lớn hơn và phức tạp hơn thì phải yêu cầu các giải pháp tự động.



Hình 1: Sơ đồ của 1 nhiệm vụ phân loại

-Ý tưởng: Yêu cầu phân loại một tập dữ liệu các trường hợp. Mỗi cá thể được đặc trưng bởi tuple(x,y). Trong đó x là tập hợp thuộc tính các giá trị mô tả (có thể chứa các tuộc tính bất thuộc loại nào), y là nhãn lớp (phải được phân loại).

-Mô hình phân loại là một đại diện trừu tượng của mối quan hẹ giữa tập thuộc tính và nhãn lớp. Chúng ta có thể biểu diễn nó một cách toán học như một hàm mục tiêu *f* dùng làm đầu vào tạo thuộc tính x và tạo ra một đầu ra tương ứng với lớp dự đoán nhãn.

- Mô hình được cho là phân loại chính xác một thể hiện (x,y) nếu *f*(x) = y.

-VD: +Lọc thư rác, x (bộ thuộc tính) là tập howpk các tính năng được trích xuất từ thông điệp email tiêu đề và nội dung, y (nhãn lớp) thư rác hoặc không phải thư rác (phân loại nhị phân).

+Phân loại thiên hà, x (bộ thuộc tính) là cacs tính năng được trích xuwts từ hình ảnh kính viễn vọng, y (nhãn lớp) hình eclip, xoẵn ốc, hoặc hình dạng không dều (phân loại đa lớp – lớn hơn 2 nhãn lớp).

-Một mô hình phân loại phục vụ 2 vai trò trong khai thác dữ liệu: predictive model (mô hình phân loại), descriptive model (mô hình miêu tả).

+Mô hình phân loại: để phân loại nững trường hợp chưa được gắn nhãn lớp trước đó.Một nô hình paan loại tốt là phải cung cấp các dự đoán chính xác với thời gian đáp ứng nhanh.

+Mô hình mô tả: để xác định các đặc điểm phân biệt các trường hợp từ các lớp khác nhau. Điều này rất có ích cho các ứng dụng quan trọng, vd: chuẩn đoán y tế, nơi không đủ có một mô hình đưa ra dự đoán mà không cần chứng minh là thế nào nó đạt được như vậy một quyết định.

-Nhãn phải là loại danh nghĩa điều này phân biệt phân loại từ các nhiệm vụ mô hình dự đoán như hồi quy, trong đó giá trị dự đoán thường là định lượng chứ không phải khi nào cũng là định tính.

-Không phải taatss cả các thuộc tính đều có liên quan đến nhiệm vụ phân loại. Vì vậy mà phát hiện sự kết hợp tối ưu của các thuộc tính phân biệt tốt nhất các trường hợp từ các lớp khác nhau là thchs thức chính trong việc xây dựng các mô hình phân loại.

-Khung phân loại chung:

+Phân loại: là nhiệm vụ gắn nhãn cho các trường hợp dữ liệu chưa được gắn nhãn.

+Tập huấn luyện: là tập dữ liệu chưa các giá trị cống nạp cũng như nhãn lớp cho từng trường hợp máy dùng để học và rút trích được những đặt điểm quan trọng để ghi nhớ lại.

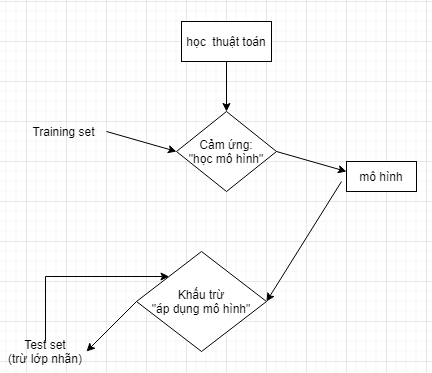
+Học thuật toán: là cách tiếp cân có hệ thống để học một mô hình phân loại được cung cấp một training set.

+Cảm ứng: là quá trình sử dụng learning algorithm để xây dựng một mô hình phân loại từ một training set.

+Khấu trừ: là quá trình áp dụng mô hình phân loại trên các trường hợp kiểm tra không nhìn thấy để dự đoán class labels.

+Qúa trình phân loại liên quan đến 2 bước: áp dụng learning algorithm để huấn luyện dữ liệu để học mô hình, áp dụng mô hình để gắn nhãn cho các trường hợp không được gắn nhãn.

+Kỹ thuật phân loại đề cập đến một cách tiếp cận chung để phân loại. Trong khi mọi mô hình định nghĩa một trình phân loại thì không phải mọi phân loại được dịnh nghĩa một mô hình duy nhất. Vd K- nearest neighbors không xây dựng một mô hình rõ ràng.



Hình 2: Khung chung để xây dựng mô hình phân loại

+Các bộ huấn luyện và kiểm tra phải độc lập với nhau để đảm bảo rằng mô hình cảm ứng có thể dự đoán chính xác các nhãn lớp trường hợp nó chưa gặp trước đây, các mô hình cung cấp dự đoán như vậy được cho là có hiệu suất khái quát tốt.

+Các lỗ mance của một mô hình (phân loại) có thể được đánh giá bằng cách so sánh dự đoán nhãn chống lại nhãn thực sự của của các trường hợp, thông tin này có thể được tóm tắt trong một bảng gọi là ma trận nhầm lẫn.

+Ma trận nhầm lân cung cấp thông tin cần thiết để xác định mô hình phân loại thực hiện tốt như thế nào, tóm tắt thông tin này thành một số duy nhất giúp thuận tiện hơn để so sánh hiệu suất tương đối của các mô hình khác nhau. Điều này được thực hiện bằng cách sử dụng một số liệu đánh giá như độ chính xác (accuracy).

(Positive là kết quả tích cực và negative là ngược lại)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted class | | |
| Actual class |  | Class = yes | Class =no |
| Class =yes | a (TP) | b (FN) |
| Class = no | c (FP) | d (TN) |

-Trong đó :

+TP: tỉ lệ các trường hợp positive đúng.

+FP: tỉ lệ các trường hợp positive sai.

+TN: tỉ lệ các trường hợp negative đúng.

+FN: tỉ lệ các trường hợp negative sai.

Accuracy = =

-Error rate (tỉ lệ lỗi) tỉ lệ dự báo sai cho tất cả trưởng hợp negative và positive

Error rate = =

Parameter :tham số truyền vào.