|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 114 | | Chương 3 | Phân loại | | | |  |  |  |  |  |  |  |
|  |  |  |  | INPUT |  |  |  |  |  | OUTPUT | | |  |  |  |
|  |  |  |  |  |  | **Mô hình  Phân loại** |  |  |  |  |  |
|  |  |  |  | Tập thuộc tính | |  |  | Nhãn | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | (**x**) |  |  |  |  |  | (*y*) | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | |  | | |  |  |  |  |  |
|  |  |  |  | **Hình 3,2.** Minh hoạ sơ đồ của một nhiệm vụ phân loại. | | | | | | | | |  |  |  |
|  |  |  | Chương này giới thiệu các khái niệm cơ bản về phân loại và mô tả | | | | | | | | | |  |  |  |
|  |  |  | một số vấn đề chính của nó như mô hình Overfitting, lựa chọn , và đánh giá mô hình | | | | | | | | | |  |  |  |
|  |  |  | . Trong khi các chủ đề được minh họa bằng cách sử dụng một kỹ thuật phân loại | | | | | | | | | |  |  |  |
|  |  |  | gọi là Cây quyết định, hầu hết các vấn đề trong chương này | | | | | | | | | |  |  |  |
|  |  |  | được áp dụng cho các kỹ thuật phân loại khác, những kỹ thuật đó sẽ được đề cập ở chương 4. | | | | | | | | | |  |  |  |
|  |  |  |  | |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 3,1 | | Khái niệm cơ bản | | |  |  |  |  |  |  |  |  |  |
|  |  |  | Hình 3,2 minh họa ý tưởng chung đằng sau sự phân loại. Dữ liệu cho một | | | | | | | | | |  |  |  |
|  |  |  | nhiệm vụ phân loại bao gồm một tập hợp các mẫu (hồ sơ). Mỗi | | | | | | | | | |  |  |  |
|  | |  | mẫu được đặc trưng bởi Tuple (x*, y*), trong đó x là tập hợp các thuộc tính | | | | | | | | | |  |  |  |
|  |  |  |  |  |
|  |  |  | giá trị mô tả mẫu và *y* là nhãn của các mẫu đó. Các | | | | | | | | | |  |  |  |
|  |  |  |  |  |
|  |  |  | thuộc tính đặt x có thể chứa các thuộc tính của bất kỳ loại nào, trong khi nhãn lớp *y* phải | | | | | | | | | |  |  |  |
|  |  |  | được phân loại hóa. | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | Một mô hình phân loại là một đại diện trừu tượng cho mối quan hệ | | | | | | | | | |  |  |  |
|  |  |  | giữa tập thuộc tính và nhãn . Sẽ được giới thiệu trong hai chương kế tiếp | | | | | | | | | |  |  |  |
|  |  |  | , mô hình có thể được đại diện bằng nhiều cách, ví dụ như một cái cây, một | | | | | | | | | |  |  |  |
|  |  |  | bảng xác suất, hoặc đơn giản là một vector có các thông số mang giá trị thực. Hình thức hơn, | | | | | | | | | |  |  |  |
|  |  |  | chúng ta có thể biểu diễn nó dưới dạng toán học như là một Target Function *f* mà phải đưa giá trị đầu vào là tập thuộc tính x và đầu ra là một nhãn được dự đoán. | | | | | | | | | |  |  |  |
|  |  |  |  | | | | | | | | | |  |  |  |
|  |  |  | Mô hình được cho là phân loại một cặp (x*, y*) chính xác nếu *f* (x) = *y*. | | | | | | | | | |  |  |  |
|  |  |  |  |  | **Bảng 3,1.** Ví dụ về các nhiệm vụ phân loại. | | | |  |  |  |  |  |  |  |
|  |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |
|  |  |  | Nhiệm vụ |  | Tập thuộc tính | |  |  |  |  | Nhãn lớp |  |  |  |  |
|  |  |  |  | |  | | | | |  |  |  |  |  |  |
|  |  |  | Lọc thư rác | | Tính năng trích xuất từ thư email | | | | |  | spam hoặc không spam |  |  |  |  |
|  |  |  |  |  | tiêu đề và nội dung | | | |  |  |  |  |  |  |  |
|  |  |  |  | |  | | | | |  |  |  |  |  |  |
|  |  |  | Nhận dạng khối u | | Tính năng chiết xuất từ ảnh scan MRI | | | | |  | ác tính hoặc lành tính |  |  |  |  |
|  |  |  |  |  |  | | | |  |  |  |  |  |  |  |
|  |  |  |  | |  | | | |  |  |  |  |  |  |  |
|  |  |  | Phân loại Thiên Hà | | Các tính năng trích xuất từ hình ảnh kính viễn vọng | | | |  |  | hình elip, xoắn ốc, hoặc |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | không đều hình |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | 3,1 | Khái niệm cơ bản | | | 115 |
|  | **Bảng 3,2.** Một dữ liệu mẫu cho vấn đề phân loại có xương sống. | | | | | |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| Vertebrate | Body | Skin | Gives | Aquatic | Aerial |  | Has |  | Hiber- |  | Class |
| Name | Temperature | Cover | Birth | Creature | Creature |  | Legs |  | nates |  | Label |
| human | warm-blooded | hair | yes | no | no |  | yes |  | no |  | mammal |
| python | cold-blooded | scales | no | no | no |  | no |  | yes |  | reptile |
| salmon | cold-blooded | scales | no | yes | no |  | no |  | no |  | fish |
| whale | warm-blooded | hair | yes | yes | no |  | no |  | no |  | mammal |
| frog | cold-blooded | none | no | semi | no |  | yes |  | yes |  | amphibian |
| komodo | cold-blooded | scales | no | no | no |  | yes |  | no |  | reptile |
| dragon |  |  |  |  |  |  |  |  |  |  |  |
| bat | warm-blooded | hair | yes | no | yes |  | yes |  | yes |  | mammal |
| pigeon | warm-blooded | feathers | no | no | yes |  | yes |  | no |  | bird |
| cat | warm-blooded | fur | yes | no | no |  | yes |  | no |  | mammal |
| leopard | cold-blooded | scales | yes | yes | no |  | no |  | no |  | fish |
| shark |  |  |  |  |  |  |  |  |  |  |  |
| turtle | cold-blooded | scales | no | semi | no |  | yes |  | no |  | reptile |
| penguin | warm-blooded | feathers | no | semi | no |  | yes |  | no |  | bird |
| porcupine | warm-blooded | quills | yes | no | no |  | yes |  | yes |  | mammal |
| eel | cold-blooded | scales | no | yes | no |  | no |  | no |  | fish |
| salamander | cold-blooded | none | no | semi | no |  | yes |  | yes |  | amphibian |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Bảng 3,1 cho thấy các ví dụ về tập thuộc tính và các nhãn cho các | | |  |  |
|  |  | nhiệm vụ phân loại. Lọc thư rác và nhận dạng khối u là ví dụ về | | |  |  |
|  |  | các vấn đề phân loại nhị phân, trong đó mỗi mẫu dữ liệu có thể được | | |  |  |
|  |  | phân loại thành một trong hai lớp. Nếu số lượng các lớp lớn hơn 2, như trong ví dụ phân loại Thiên Hà | | |  |  |
|  |  | ,Đó được gọi là phân loại đa lớp. | | |  |  |
|  |  | Chúng ta minh họa các khái niệm cơ bản về phân loại trong chương này với | | |  |  |
|  |  | hai ví dụ sau. | | |  |  |
|  |  | Ví dụ 3,1. [Phân loại động vật có xương sống] Bảng 3,2 Hiển thị dữ liệu mẫu | | |  |  |
|  |  | phân loại động vật có xương sống thành động vật có vú, bò sát, chim, cá, và | | |  |  |
|  |  | lưỡng cư. Tập thuộc tính bao gồm các đặc điểm của vật có xương sống như | | |  |  |
|  |  | nhiệt độ cơ thể, lớp da và khả năng bay. Tập dữ liệu cũng có thể được | | |  |  |
|  |  | sử dụng cho một nhiệm vụ phân loại nhị phân như phân loại động vật có vú, bằng cách nhóm | | |  |  |
|  |  | Các loài bò sát, chim, cá, và động vật lưỡng cư thành một loại duy nhất được gọi là | | |  |  |
|  |  | động vật không có vú. |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  | Ví dụ 3,2. [ Phân loại vay mượn] Xem xét các vấn đề của | | |  |  |
|  |  | dự đoán là một bên vay mượn sẽ trả nợ hoặc không các khoản vay. | | |  |  |
|  |  | Tập dữ liệu được sử dụng để xây dựng mô hình phân loại được hiển thị trong bảng | | |  |  |
|  |  | 3,3. Tập thuộc tính bao gồm thông tin cá nhân của bên vay như | | |  |  |
|  |  | tình trạng hôn nhân và thu nhập hàng năm, trong khi các nhãn cho biết liệu | | |  |  |
|  |  | người vay đã vỡ nợ và không thể chi trả khoản vay. |  |  |  |  |
|  |  |  |  |  |

116 chương 3 phân loại

**Bảng 3,3.** Một dữ liệu mẫu cho các vấn đề phân loại người vay.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Home Owner | Marital Status | Annual Income | Defaulted? |
|  |  |  |  |  |
| 1 | Yes | Single | 125000 | No |
|  |  |  |  |  |
| 2 | No | Married | 100000 | No |
|  |  |  |  |  |
| 3 | No | Single | 70000 | No |
|  |  |  |  |  |
| 4 | Yes | Married | 120000 | No |
|  |  |  |  |  |
| 5 | No | Divorced | 95000 | Yes |
|  |  |  |  |  |
| 6 | No | Single | 60000 | No |
|  |  |  |  |  |
| 7 | Yes | Divorced | 220000 | No |
|  |  |  |  |  |
| 8 | No | Single | 85000 | Yes |
|  |  |  |  |  |
| 9 | No | Married | 75000 | No |
|  |  |  |  |  |
| 10 | No | Single | 90000 | Yes |
|  |  |  |  |  |

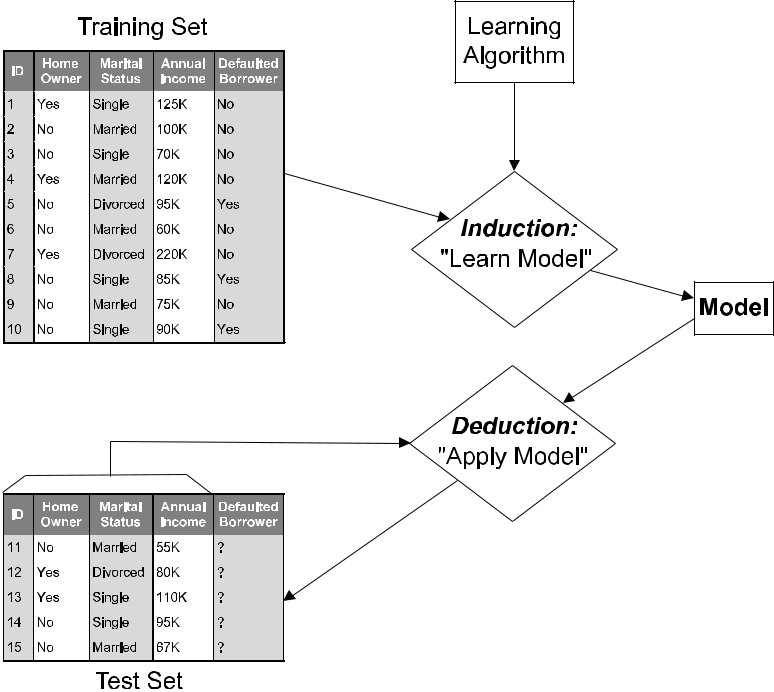
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | Một mô hình phân loại phục vụ hai vai trò quan trọng trong khai thác dữ liệu. Đầu tiên, nó được | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | sử dụng như một mô hình dự đoán để phân loại các mẫu chưa được gắn nhãn. Một | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | mô hình phân loại tốt phải cung cấp dự đoán chính xác với một tốc độ phản hồi nhanh chóng. | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | Thứ hai, nó phục vụ như một mô hình mô tả để xác định các đặc điểm | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | phân biệt các trường hợp từ các lớp khác nhau. Điều này đặc biệt hữu ích | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | cho các ứng dụng quan trọng, chẳng hạn như chẩn đoán y tế, nơi không đủ | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | có một mô hình mà làm cho một dự đoán mà không đủ biện minh về quyết định của nó | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | Ví dụ, một mô hình phân loại sinh ra từ bộ dữ liệu động vật có xương sống | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | Được minh họa trong bảng 3,2 có thể được sử dụng để dự đoán nhãn lớp của các | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | động vật có xương sống. | | | |  | | | |  |  |  |  | | |  |  |  |  |  |  |  |  |
|  |  | | | | Ngoài ra, nó có thể được sử dụng như một mô hình miêu tả để giúp xác định các loài động vật | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | lưỡng cư. Ví dụ, mô hình có thể xác định động vật có vú là động vật máu nóng, | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | có xương sống và sinh con. | | | | | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |
|  |  | | | | Có một số điểm đáng chú ý liên quan đến ví dụ trước. | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | Đầu tiên, mặc dù tất cả các thuộc tính được hiển thị trong bảng 3,2 là rõ ràng, | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | không có giới hạn về loại thuộc tính có thể được sử dụng như các biến dự đoán. | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | Mặt khác, các nhãn phải có nghĩa. Tức phân biệt phân loại với cac nhiệm vụ mô | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | phân loại với các nhiệm vụ mô hình dự đoán khác như hồi quy, nơi | | | | | | | | | | | | | | | | | | | | | |
|  |  | | | | giá trị dự đoán thường được định lượng. | | | | | | | | | | | | | | | | | | | | | |
|  |  |  |  |  | |  |  |  | |  |  |  | | | | |  |

3,2 117

3.2 Framework chung cho sự phân loại

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Phân loại là nhiệm vụ gán nhãn cho các mẫu dữ liệu không được gắn nhãn và một |  |  |  |
|  |  |  | các phân loại được sử dụng để thực hiện một nhiệm vụ như vậy. Một loại phân loại thường được mô tả trong |  |  |  |
|  |  |  | điều khoản của một mô hình như minh họa trong phần trước. Mô hình được tạo ra |  |  |  |
|  |  |  | sử dụng một tập hợp các mẫu cho trước, được gọi là tập Training, trong đó chứa |  |  |  |
|  |  |  | giá trị thuộc tính cũng như nhãn cho mỗi trường hợp. Cách tiếp cận có hệ thống |  |  |  |
|  |  |  |  |
|  |  | cho một mô hình học phân loại một tập Training được biết đến như là một |  |  |
|  |  |  | Thuật toán học. Quá trình sử dụng một thuật toán học để xây dựng một phân loại |  |  |  |
|  |  |  | Mô hình từ dữ liệu đào tạo được gọi là cảm ứng. Quá trình này cũng |  |  |  |
|  |  |  | thường được mô tả là "học một model" hoặc "xây dựng một model." Quá trình này |  |  |  |
|  |  |  | áp dụng một mô hình phân loại trên các phiên bản thử nghiệm vô thần để dự đoán lớp học của |  |  |  |
|  |  |  | nhãn được gọi là khấu trừ. Như vậy, quá trình phân loại bao gồm hai |  |  |  |
|  |  |  | Các bước: áp dụng một thuật toán học tập để đào tạo dữ liệu để học một mô hình, và |  |  |  |
|  |  |  | sau đó áp dụng mô hình để gán nhãn cho các trường hợp không có nhãn. Hình 3,3 |  |  |  |
|  |  |  | minh họa khuôn khổ chung để phân loại. |  |  |  |
|  |  |  | Một kỹ thuật phân loại đề cập đến một cách tiếp cận chung để phân loại, |  |  |  |
|  |  |  | Ví dụ như, kỹ thuật cây quyết định mà chúng ta sẽ học trong chương này. Này |  |  |  |
|  |  |  | kỹ thuật phân loại như hầu hết những người khác, bao gồm một gia đình của các mô hình liên quan |  |  |  |
|  |  |  | và một số thuật toán để học các mô hình này. Trong chương 4, chúng ta |  |  |  |
|  |  |  | sẽ học các kỹ thuật phân loại bổ sung, bao gồm mạng nơ-ron và |  |  |  |
|  |  |  | hỗ trợ máy vector. |  |  |  |
|  |  |  | Một vài ghi chú về thuật ngữ. Đầu tiên, các thuật ngữ "classifier" và "mô hình" |  |  |  |
|  |  |  | thường được lấy là đồng nghĩa. Nếu một kỹ thuật phân loại xây dựng một đơn, |  |  |  |
|  |  |  | Mô hình toàn cầu, thì điều này là tốt. Tuy nhiên, trong khi mọi mô hình xác định một classifier, |  |  |  |
|  |  |  | không phải tất cả các phân loại được xác định bởi một mô hình duy nhất. Một số loại phân loại, như *k*- |  |  |  |
|  |  |  | hàng xóm gần nhất, không xây dựng một mô hình rõ ràng (mục 4,3), trong khi |  |  |  |

118 chương 3 phân loại



**Hình 3,3.** Khuôn khổ chung để xây dựng một mô hình phân loại.

Các phân loại khác, chẳng hạn như các loại phân loại, kết hợp đầu ra của một bộ sưu tập các mô hình (mục 4,10). Thứ hai, thuật ngữ "classifier" thường được sử dụng trong một ý nghĩa chung hơn để đề cập đến một kỹ thuật phân loại. Vì vậy, ví dụ, "phân loại cây quyết định" có thể tham khảo các kỹ thuật loại cây quyết định hoặc một phân loại cụ thể được xây dựng bằng cách sử dụng kỹ thuật đó. May mắn thay, ý nghĩa của "classifier" thường là rõ ràng từ ngữ cảnh.

Trong khuôn khổ chung Hiển thị trong hình 3,3, các bước cảm ứng và khấu trừ sẽ được thực hiện riêng. Trong thực tế, như sẽ được thảo luận sau này trong phần 3,6, các tập huấn và thử nghiệm phải được độc lập với nhau để đảm bảo rằng mô hình gây ra có thể dự đoán chính xác các nhãn lớp của trường hợp nó chưa bao giờ gặp phải trước đó. Mô hình cung cấp thông tin chi tiết tiên đoán như vậy được cho là có hiệu suất tổng quát tốt. Các Perfor-Mance của một mô hình (classifier) có thể được đánh giá bằng cách so sánh các nhãn dự đoán chống lại các nhãn thực sự của các trường hợp. Thông tin này có thể được tóm tắt trong một bảng gọi là một ma trận nhầm lẫn. Bảng 3,4 mô tả các ma trận nhầm lẫn cho một vấn đề phân loại nhị phân. Mỗi mục *fIJ là* số lượng các phiên bản từ lớp *tôi* dự đoán sẽ thuộc lớp *j*. Ví dụ: *f*01 là

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  | 3,3 |  | Quyết định cây Classifier | | | | | | 119 |  |  |  |
|  |  |  | **Bảng 3,4.** Nhầm lẫn ma trận cho một vấn đề phân loại nhị phân. | | | | | | | | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | Dự đoán Class | | | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | Lớp học = 1 | |  | Lớp học = 0 | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | Thực tế | Lớp học = 1 | | | *f*11 | |  | *f*10 | | |  |  |  |  |  |  |
|  |  |  |  |  | Lớp | Lớp học = 0 | | | *f*01 | |  | *f*00 | | |  |  |  |  |  |  |
|  |  |  | số trường hợp từ lớp 0 không chính xác dự đoán là lớp 1. Số lượng | | | | | | | | | | | | | | | | | |
|  |  |  | Các dự đoán chính xác được thực hiện bởi mô hình là (*f*11 + *f*00) và số lượng | | | | | | | | | | | | | | | | | |
|  |  |  | dự đoán không chính xác là (*f*10 + *f*01). | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | Mặc dù ma trận nhầm lẫn cung cấp thông tin cần thiết để xác định | | | | | | | | | | | | | | | | | |
|  |  |  | một mô hình phân loại thực hiện tốt như thế nào, tóm tắt thông tin này vào một | | | | | | | | | | | | | | | | | |
|  |  |  | số duy nhất làm cho nó thuận tiện hơn để so sánh hiệu suất tương đối | | | | | | | | | | | | | | | | | |
|  |  |  | của các mô hình khác nhau. Điều này có thể được thực hiện bằng cách sử dụng một số liệu đánh giá như | | | | | | | | | | | | | | | | | |
|  |  |  | độ chính xác, được tính theo cách sau: | | | | | | | | | | | | |  |  |  |  |  |
|  |  |  |  | Độ chính xác = | | | Số dự đoán chính xác | | | | | | | *.* | |  | (3,1) |  |  |  |
|  |  |  |  |  | |  | |  |  |  |  |  |  |
|  |  |  |  |  |  |  | Tổng số dự đoán | | | | | | | | |  |  |  |  |  |
|  |  |  | Đối với vấn đề phân loại nhị phân, độ chính xác của một mô hình được cho bởi | | | | | | | | | | | | | |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | Độ chính xác = | | |  | *f*11 +  *f*00 | | | | | | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | |  | |  | *.* | | |  | (3,2) |  |  |  |
|  |  |  |  |  | *f*11 +  *f*10 +  *f*01 +  *f*00 | | | | |  |  |  |  |
|  |  |  | Tỷ lệ lỗi là một số liệu liên quan, được định nghĩa như sau cho nhị phân | | | | | | | | | | | | | | | | | |
|  |  |  | vấn đề phân loại: | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | Tỷ lệ lỗi = | Số dự đoán sai | | | | | | = |  | *f*10 +  *f*01 | | | | *.* | (3,3) |  |  |  |
|  |  |  |  | | | | | |  | *f*11 +  *f*10 +  *f*01 +  *f*00 | | | |  |  |  |
|  |  |  |  | Tổng số dự đoán | | | | | | |  |  |  |  |  |  |

Các thuật toán học tập của hầu hết các kỹ thuật phân loại được thiết kế để các mô hình học đạt được độ chính xác cao nhất, hoặc tương đương, tỷ lệ lỗi thấp nhất khi áp dụng cho bộ kiểm tra. Chúng tôi sẽ xem xét lại chủ đề của mô hình đánh giá trong phần 3,6.

3.3 Kiểu phân loại cây quyết định : Phần này giới thiệu một kỹ thuật phân loại đơn giản được biết đến là “Cây quyết định” . Để minh họa cách một cây quyết định hoạt động, hãy cân nhắc vấn đề phân loại của động vật không có vú sử dụng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | 120 | Chương 3 phân loại |  |  |
|  |  | tập dữ liệu động vật có xương sống được minh họa trong Bảng 3.2. Giả sử một loài mới được phát hiện | |  |  |
|  |  | bởi các nhà khoa học. Làm thế nào chúng ta nói đó có phải là một động vật có vú hay không ? Một | |  |  |
|  |  | cách tiếp cận là đặt ra một loạt các câu hỏi về các đặc điểm của loài này. | |  |  |
|  |  | Câu hỏi đầu tiên liệu đó là loài máu lạnh hay máu nóng . | |  |  |
|  |  | Nếu nó là máu lạnh thì chắc chắn không phải là một động vật có vú. Nếu không, nó là một trong hai | |  |  |
|  |  | Hoặc là chim hoặc là động vật có vú. Trường hợp thứ hai, sinh con hay không? Nếu sinh con chắc chắn đó là động vật có vú( | |  |  |
|  |  |  | |  |  |
|  |  | sinh là động vật có vú chắc chắn, trong khi những người không có khả năng là không | |  |  |
|  |  | động vậtcó vú (ngoại trừ các động vật có vú đẻ trứng như thú mỏ vịt | |  |  |
|  |  | và gai anteater). | |  |  |
|  |  | Ví dụ trước minh hoạ cách chúng ta có thể giải quyết một vấn đề phân loại | |  |  |
|  |  | bằng cách hỏi một loạt các câu hỏi cẩn thận về các thuộc tính của bài kiểm tra | |  |  |
|  |  | Ví dụ. Mỗi lần chúng tôi nhận được một câu trả lời, chúng tôi có thể yêu cầu một câu hỏi theo dõi | |  |  |
|  |  | cho đến khi chúng tôi có thể conclusively quyết định trên nhãn lớp của nó. Loạt các câu hỏi và | |  |  |
|  |  | có thể được tổ chức thành một cấu trúc phân cấp được gọi là một | |  |  |
|  |  | cây quyết định. Hình 3,4 cho thấy một ví dụ về cây quyết định cho các động vật có vú | |  |  |
|  |  | vấn đề phân loại. Cây có ba loại nút: | |  |  |
|  |  | *•* Một nút gốc, không có liên kết đến và không hoặc nhiều liên kết đi. | |  |  |
|  |  | *•* | Các nút nội bộ, mỗi trong số đó có chính xác một liên kết đến và hai |  |  |
|  |  |  |  |

hoặc nhiều liên kết đi.

*•* Lá hoặc nút thiết bị đầu cuối , mỗi trong số đó có chính xác một liên kết đến và không có liên kết đi.

Mỗi nút lá trong cây quyết định được liên kết với một nhãn lớp. Các nút không phải đầu cuối, bao gồm các nút gốc và nội bộ, chứa các điều kiện kiểm tra thuộc tính thường được xác định bằng một thuộc tính duy nhất. Mỗi kết quả có thể xảy ra trong điều kiện kiểm tra thuộc tính liên quan đến chính xác một con của nút này. Ví dụ: nút gốc của cây Hiển thị trong hình 3,4 sử dụng thuộc tính cơ thể nhiệt độ để xác định một điều kiện kiểm tra thuộc tính có hai kết quả, ấm áp và lạnh, dẫn đến hai nút con.

Đưa ra một cây quyết định, phân loại một phiên bản thử nghiệm là đơn giản. Bắt đầu từ nút gốc, chúng tôi áp dụng điều kiện kiểm tra thuộc tính của nó và làm theo chi nhánh phù hợp dựa trên kết quả kiểm tra. Điều này sẽ dẫn chúng ta đến một nút nội bộ khác, mà một tình huống kiểm tra thuộc tính mới được áp dụng hoặc một nút lá. Khi đạt đến nút lá, chúng tôi chỉ định nhãn lớp được liên kết với nút này cho phiên bản thử nghiệm. Như một minh hoạ, hình 3,5 dấu vết đường dẫn được sử dụng để dự đoán nhãn lớp của một Flamingo. Đường dẫn chấm dứt ở một nút lá có nhãn là không có động vật có vú.

Kh