**BÀI TẬP KHAI PHÁ DỮ LIỆU**

**HỌ TÊN: HOÀNG THỊ HOÀ**

**MSV: 637634**

# Tổng quan

1. Khai phá dữ liệu là gì? Các chức năng của khai phá dữ liệu?

Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó

Các chức năng của khai phá dữ liệu là:

* Mô tả khái niệm: thiên về mô tả, tổng hợp tóm tắt khái niệm. Ví dụ tóm tắt văn bản
* Luật kết hợp: là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Ví dụ: 50% nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin-sinh, tài chính & thị trường chứng khoán…
* Phân lớp và dự đoán: xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định, mạng nơ ron nhân tạo … Người ta còn gọi phân lớp là học có giám sát
* Phân cụm: xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát
* Khai phá chuỗi: tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao

1. Hãy lấy một ví dụ thực tế mà ở đó việc khai phá dữ liệu là vấn đề quan trọng cho sự thành công của hoạt động của tổ chức/doanh nghiệp. Hãy làm rõ:

* Các nguồn dữ liệu, các loại dữ liệu có thể có của tổ chức/ doanh nghiệp - Các bài toán khai phá dữ liệu và tập dữ liệu liên quan dùng để khai phá.
* Các kỹ thuật khai phá dữ liệu?
* Các công cụ cho khai phá dữ liệu?
* Kết quả của khai phá dữ liệu và khả năng ứng dụng?

Ví dụ: việc khai phá dữ liệu với google và facebook.

Nguồn dữ liệu họ thu thập trên lịch sử duyệt web, những từ khoá bạn tìm kiếm, ngôn ngữ bạn đang dùng, mạng xã hội của bạn cũng như vị trí hiện tại, các liên hệ cá nhân, các cuộc hội thoại, nội dung trao đổi, những dữ liệu bạn đăng tải lên

1. Tại sao có các phần tử ngoại lai (outliers) trong dữ liệu? Cách phát hiện chúng?

Outliers: những dữ liệu (đối tượng) không tuân theo đặc tính/ hành vi chung của tập dữ liệu (đối tượng)

Nguyên nhân có các phần tử ngoại lai trong dữ liệu là:

* Khác quan: công cụ thu thập dữ liệu, lỗi trên đường truyền, giới hạn công nghệ,…
* Chủ quan: do con người

Cách phát hiện chúng:

* Dựa trên phân bố thống kê
* Dựa trên khoảng cách
* Dựa trên mật độ
* Dựa trên độ lệch

1. Tại sao phải làm sạch dữ liệu?

Tại vì:

* Nếu dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn) thì các kết quả khai phá dữ liệu sẽ bị ảnh hưởng và không đáng tin cậy
* Các kết quả khai phá dữ liệu (các tri thức khám phá được) không chính xác (không đáng tin cậy) sẽ dẫn đến các quyết định không chính, không tối ưu
* Ví dụ: Các dữ liệu chứa lỗi hoặc thiếu giá trị thuộc tính sẽ có thể dẫn đến các kết quả thống kê sai lầm

# Dữ liệu và tiền xử lý dữ liệu

1. Bài tập số 2.2 trang 80

Giả sử rằng dữ liệu để phân tích bao gồm thuộc tính tuổi. Giá trị độ tuổi cho tập dữ liệu đang tăng dần

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30,

33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Tính mean? Tính median?

(b) Tính mode? Nhận xét về phương thức của dữ liệu(hai phương thức, ba phương thức…)

(c) Tính midrange?

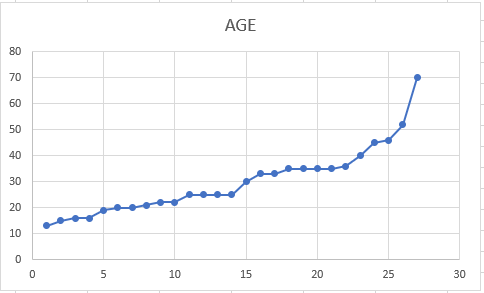
(d) Bạn có thể tìm thấy (roughly/gần) tứ phân vị đầu tiên (Q1) và tứ phân vị thứ ba (Q3) của dữ liệu không?

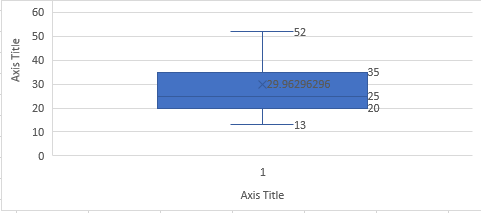
(e) Cung cấp five-number tóm tắt của dữ liệu

(f) Hiển thị boxplot của dữ liệu.

(g) Lượng tử-lượng tử khác với biểu đồ lượng tử như thế nào?

Bài làm:

1. Mean = 29.963; median = 25; std = 12.942
2. Mode = 25 xuất hiện 4 lần
3. Min = 13; Max = 70; Midrange = 41,5
4. Ta tìm thấy tứ phân vị đầu tiên = 20.5; tứ phân vị thứ ba ko tìm thấy được
5. Mean = 29.963; median = 25; std = 12.942; min = 13; max = 70
6. 



1. Bài tập số 2.4 trang 80.

Giả sử bệnh viện kiểm tra dữ liệu tuổi và mỡ cơ thể cho 18 người lớn được chọn ngẫu nhiên

Với các kết quả sau:

Table

Description automatically generated

1. Tính mean, median, and standard deviation của tuổi and %fat.

(Tính trung bình, trung vị và độ lệch chuẩn của tuổi và %fat.)

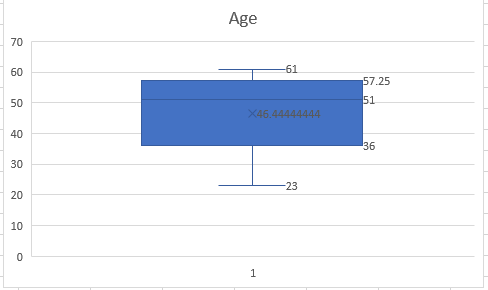
(b) Vẽ boxplots cho tuổi và %fat

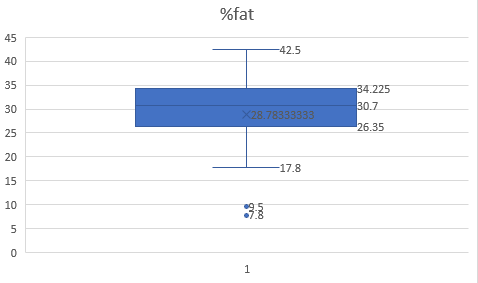
(c) Vẽ một biểu đồ tán xạ và một biểu đồ q-q dựa trên hai biến này

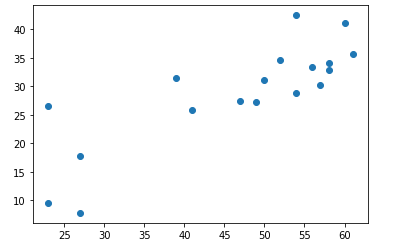
Bài làm

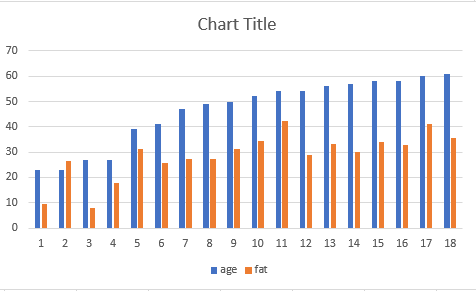
1. Tuổi: mean = 46.444, median = 51, std = 13.219

%fat: mean = 28.783, median = 30.7 , std = 9.254

1. 



1. 



1. Bài tập số 2.8 trang 81.

Điều quan trọng là phải xác định hoặc chọn độ đo tương tự trong phân tích dữ liệu. Tuy nhiên ở đó không có độ đo tương tự phổ biến thường được chấp nhận. Kết quả có thể thay đổi tuỳ thuộc vào độ đo tương tự được sử dụng. Tuy nhiên, các độ đo tương đồng dường như khác nhau có thể tương đương sau một số biến đổi.

Giả sử chúng ta có tập dữ liệu 2-D sau:

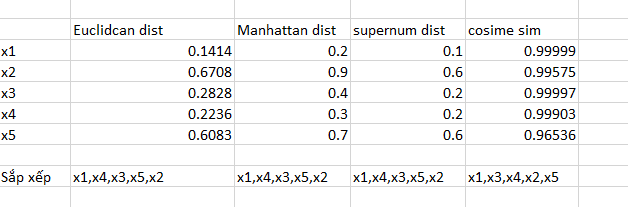
Table

Description automatically generated

(a) Coi dữ liệu là điểm dữ liệu 2-D. Với một điểm dữ liệu mới, x = (1,4,1,6) dưới dạng truy vấn, xếp hạng các điểm cơ sở dữ liệu dựa trên sự tương đồng với truy vấn bằng cách sử dụng Euclidean khoảng cách Manhattan, khoảng cách xa nhất và sự tương đồng cosin

(b) Chuẩn hóa tập dữ liệu để làm cho định mức của mỗi điểm dữ liệu bằng 1. Sử dụng Euclide khoảng cách trên dữ liệu được chuyển đổi để xếp hạng các điểm dữ liệu.

Bài làm



1. Chuẩn hoá

Trang tính

1. Bài tập số 3.2 trang 121

Trong dữ liệu trong thế giới thực, các tuples bị thiếu giá trị cho một số thuộc tính là xuất hiện phổ biến. Mô tả các phương pháp khác nhau để xử lý vấn đề này.

Các phương pháp khác nhau để xử lý vấn đề thiếu giá trị trong bộ dữ liệu bao gồm:

(a) Bỏ qua tuple: Điều này thường được thực hiện khi giá trị bị thiếu. Phương pháp này không hiệu quả lắm trừ khi bộ tuple chứa một số thuộc tính bị thiếu giá trị. Nó đặc biệt kém khi tỷ lệ phần trăm giá trị bị thiếu trên mỗi thuộc tính thay đổi đáng kể.

(b) Điền giá trị còn thiếu theo cách thủ công: Nói chung, cách tiếp cận này tốn thời gian và có thể không phải là nhiệm vụ hợp lý đối với các tập dữ liệu lớn có nhiều giá trị bị thiếu, đặc biệt là khi giá trị cần điền

không dễ dàng xác định.

(c) Sử dụng hằng số toàn cục để điền vào giá trị bị thiếu: Thay thế tất cả các giá trị thuộc tính bị thiếu bằng cùng một hằng số, chẳng hạn như nhãn như *\ Unknown, "* Do đó, mặc dù phương pháp này đơn giản nhưng không được khuyến khích.

(d) Sử dụng thuộc tính trung bình cho các giá trị định lượng (số) hoặc chế độ thuộc tính cho các giá trị phân loại (danh nghĩa): Ví dụ: giả sử rằng thu nhập trung bình của khách hàng *AllElectronics* là 28.000 đô la. Sử dụng giá trị này để thay thế bất kỳ giá trị nào còn thiếu cho *thu nhập* .

(e) Sử dụng giá trị có thể xảy ra nhất để điền vào giá trị còn thiếu

1. Bài tập số 3.3 trang 121

Bài tập 2.2 đã đưa ra các dữ liệu sau (theo thứ tự tăng) cho độ tuổi thuộc tính: 13, 15,

16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46,

52, 70.

(a) Use smoothing by bin có nghĩa là làm sạch các dữ liệu này, using a bin depth of 3. Minh hoạ các bước của bạn. Nhận xét về hiệu quả của kỹ thuật này đối với dữ liệu nhất định.

(b) Làm thế nào bạn có thể xác định ngoại lệ/ ngoại lai trong dữ liệu?

(c) Các phương pháp làm sạch dữ liệu?

Bài làm:

1. Nhận xét

Kỹ thuật này được sử dụng để làm mịn dữ liệu. Dữ liệu có thể có nhiễu, bằng cách sử dụng kỹ thuật binning, chúng tôi có thể loại bỏ nhiễu khỏi dữ liệu. Nó giúp cung cấp kết quả chính xác hơn

Các yếu tố ngoại lai trong dữ liệu có thể được xác định theo một số cách.

- Bằng cách chia dữ liệu thành các biểu đồ có chiều rộng tương đương và xác định phần bên ngoài

biểu đồ

- Bằng cách phân cụm dữ liệu thành các nhóm. Bất kỳ dữ liệu nào không thuộc nhóm có thể được coi là

những ngoại lệ.

- Nói chung, phù hợp với một mô hình với dữ liệu. Bất kỳ điểm dữ liệu nào sai lệch đáng kể (dựa trên

một số ngưỡng) từ mô hình có thể được coi là ngoại lệ.

1. Bài tập số 3.7 trang 121

Sử dụng dữ liệu cho độ tuổi được đưa ra trong Bài tập 3.3, hãy trả lời như sau:

(a) Sử dụng chuẩn hóa tối đa tối thiểu để chuyển đổi giá trị 35 cho độ tuổi vào phạm vi [0.0,1.0].

(b) Sử dụng chuẩn hóa điểm z để chuyển đổi giá trị 35 cho độ tuổi, trong đó tiêu chuẩn độ tuổi là 12,94 tuổi.

(c) Sử dụng chuẩn hóa theo tỷ lệ thập phân để chuyển đổi giá trị 35 cho độ tuổi.

(d) Nhận xét về phương pháp bạn muốn sử dụng cho dữ liệu nhất định, đưa ra lý do tại sao

1. Bài tập số 3.9 trang 122.

Giả sử một nhóm gồm 12 bản ghi giá bán đã được sắp xếp như sau:

5,10,11,13,15,35,50,55,72,92,204,215.

Phân vùng chúng thành ba thùng theo từng phương pháp sau:

(a) phân vùng tần số bằng nhau (độ sâu bằng nhau)

(b) phân vùng có chiều rộng bằng nhau

(c) phân cụm

# Các kỹ thuật phân lớp

1. Bài 8.1 trang 386

Phác thảo ngắn gọn các bước chính của phân loại cây quyết định

Các bước chính như sau:

• Cây bắt đầu như một nút gốc duy nhất chứa tất cả các bộ đào tạo.

• Nếu các bộ giá trị đều thuộc cùng một lớp, thì nút sẽ trở thành một lá, được gắn nhãn với lớp đó.

• Ngoài ra, một phương pháp lựa chọn thuộc tính được gọi để xác định tiêu chí tách. Một phương pháp như vậy

có thể sử dụng một phép đo kinh nghiệm hoặc thống kê (ví dụ: mức tăng thông tin, tỷ lệ khuếch đại hoặc chỉ số gini) để

chọn cách "tốt nhất" để tách các bộ giá trị thành các lớp riêng lẻ. Tiêu chí phân tách bao gồm thuộc tính phân tách và cũng có thể chỉ ra điểm phân tách hoặc tập hợp con phân tách, như được mô tả

phía dưới.

• Tiếp theo, nút được gắn nhãn với tiêu chí phân tách, dùng như một phép thử tại nút. A nhánh được phát triển từ nút đến từng kết quả của tiêu chí phân tách và các bộ giá trị được phân vùng cho phù hợp. Có ba trường hợp có thể xảy ra cho việc phân vùng như vậy. (1) Nếu thuộc tính tách có giá trị rời rạc, sau đó một nhánh được phát triển cho mỗi giá trị có thể có của thuộc tính.

(2) Nếu thuộc tính tách, A, có giá trị liên tục, thì hai nhánh được phát triển, tương ứng với các điều kiện A ≤ điểm phân chia và A> điểm phân chia. (3) Nếu thuộc tính phân tách có giá trị rời rạc và phải tạo cây nhị phân (ví dụ: nếu chỉ mục gini được sử dụng làm thước đo lựa chọn),

thì kiểm tra tại nút là "A ∈ SA?" trong đó SA là tập con tách cho A. Nó là tập con của các giá trị đã biết của A. Nếu một bộ giá trị đã cho có giá trị aj của A và nếu aj ∈ SA, thì kiểm tra tại nút được thỏa mãn.

• Thuật toán đệ quy để tạo cây quyết định cho các bộ giá trị tại mỗi phân vùng.

Các điều kiện dừng là:

• Nếu tất cả các bộ giá trị tại một nút nhất định thuộc cùng một lớp, thì hãy biến đổi nút đó thành một lá, có nhãn với lớp đó.

• Nếu không còn thuộc tính nào nữa để tạo thêm phân vùng, thì có thể sử dụng biểu quyết đa số để

chuyển đổi nút đã cho thành một lá, được gắn nhãn với lớp phổ biến nhất trong số các bộ giá trị.

• Nếu không có bộ giá trị nào cho một nhánh nhất định, một lá sẽ được tạo với lớp đa số từ cấp chính nút

1. Bài 8.6 trang 387

Tại sao phân loại naive Bayesian classification called “naive”? Phác thảo các ý tưởng chính của na¨ıve Bayesian classification.

Phân loại theo kiểu Bayes chưa được gọi là na¨ıve vì nó giả định tính độc lập có điều kiện của lớp.

Có nghĩa là, ảnh hưởng của một giá trị thuộc tính trên một lớp nhất định là độc lập với các giá trị của các thuộc tính khác. Giả định này được thực hiện để giảm chi phí tính toán, và do đó được coi là "không có". Các ý tưởng chính đằng sau phân loại Bayes là cố gắng và phân loại dữ liệu bằng cách tối đa hóa P (X | Ci) P (Ci)

(trong đó i là một chỉ số của lớp) bằng cách sử dụng định lý Bayes về xác suất hậu nghiệm. Nói chung:

• Chúng tôi được cung cấp một tập hợp các bộ dữ liệu không xác định, trong đó mỗi bộ được biểu diễn bằng một n-chiều

vector, X = (x1, x2,.., xn) mô tả n phép đo được thực hiện trên bộ từ n thuộc tính,

lần lượt là A1, A2, .., An. Chúng ta cũng được cung cấp một tập hợp m lớp, C1, C2,. . . Cm.

• Sử dụng định lý Bayes, bộ phân loại Bayes bình thường tính xác suất sau của mỗi

lớp có điều kiện trên X. X được gán nhãn lớp của lớp có giá trị đứng sau lớn nhất xác suất điều kiện trên X. Do đó, chúng tôi cố gắng tối đa hóa P (Ci | X) = P (X | Ci) P (Ci) / P (X).

Tuy nhiên, vì P (X) là hằng số đối với tất cả các lớp nên chỉ cần P (X | Ci) P (Ci) là cực đại. Nếu xác suất trước của lớp không được biết, sau đó người ta thường cho rằng các lớp là như nhau

có thể, tức là P (C1) = P (C2) = · · · = P (Cm), và do đó chúng ta sẽ tối đa hóa P (X | Ci). Nếu không, chúng ta tối đa hóa P (X | Ci) P (Ci). Xác suất trước của lớp có thể được ước tính bằng P (Ci) = si

S, si ở đâu là số lượng các bộ đào tạo của lớp Ci, và s là tổng số bộ giá trị đào tạo.

• Để giảm tính toán trong việc đánh giá P (X | Ci), giả thiết cơ bản về tính độc lập có điều kiện của lớp được đưa ra. Điều này giả định rằng các giá trị của các thuộc tính có điều kiện

độc lập với nhau, với nhãn lớp của bộ tuple, tức là không có sự phụ thuộc mối quan hệ giữa các thuộc tính.

- Nếu Ak là thuộc tính phân loại thì P (xk | Ci) bằng số bộ giá trị đào tạo trong Ci có xk là giá trị cho thuộc tính đó, chia cho tổng số bộ giá trị đào tạo ở Ci.

- Nếu Ak là thuộc tính liên tục thì P (xk | Ci) có thể được tính bằng mật độ Gaussian chức năng.

1. Bài 8.7 trang 387

Bảng sau đây bao gồm dữ liệu đào tạo từ cơ sở dữ liệu nhân viên. Dữ liệu có đã được khái quát hóa. Ví dụ: "31 ... 35" cho độ tuổi đại diện cho độ tuổi từ 31 đến 35

Đối với một mục nhập hàng nhất định, số count đại diện cho số lượng tuple dữ liệu có giá trị cho department, status, age, and salary đưa ra trong hàng đó

Graphical user interface, table

Description automatically generated

Let status be the class label attribute.

(a) Làm thế nào bạn sẽ sửa đổi thuật toán cây quyết định cơ bản cần xem xét số lượng của mỗi tuple dữ liệu tổng quát (tức là của mỗi mục hàng)?

(b) Sử dụng thuật toán của bạn để xây dựng một cây quyết định từ dữ liệu đã cho.

(c) Đưa ra một tuple dữ liệu có các giá trị "system", "26 . . . 30," và "46–50K" cho thuộc tính bộ phận, tuổi tác và tiền lương, tương ứng, một na¨ıve Bayesian phân loại tình trạng cho tuple được?

1. Bài 8.14, 8.15, 8.16

**8.14**

Giả sử rằng chúng ta muốn chọn giữa hai mô hình dự đoán, M1 và M2. Chúng tôi có

thực hiện 10 vòng xác thực chéo gấp 10 lần trên mỗi mô hình, trong đó cùng một dữ liệu

phân vùng trong vòng i được sử dụng cho cả M1 và M2. Tỷ lệ lỗi thu được cho M1 là 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0. Tỷ lệ lỗi cho M2 là 22.4,14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. Nhận xét về việc một mô hình có phải là

tốt hơn đáng kể so với mức khác xem xét mức độ quan trọng là 1%.

**8.15**

What is boosting? Nêu rõ lý do tại sao nó có thể cải thiện độ chính xác của cảm ứng cây quyết định

Table

Description automatically generated

Bootsting là một kỹ thuật được sử dụng để giúp cải thiện độ chính xác của bộ phân loại. Chúng ta được cung cấp một tập S gồm các bộ giá trị. Đối với lần lặp t, trong đó t = 1, 2 ,. . . , T, tập huấn luyện St được lấy mẫu thay thế từ S. Gán trọng số vào các bộ giá trị trong tập huấn luyện đó. Tạo bộ phân loại, Ct từ St. Sau khi Ct được tạo, hãy cập nhật trọng số của các bộ giá trị để các bộ giá gây ra lỗi phân loại sẽ có xác suất lớn hơn đang được chọn để xây dựng bộ phân loại tiếp theo. Bộ phân loại boosting cuối cùng kết hợp các lần chọn của từng bộ phân loại riêng lẻ, trong đó trọng số của mỗi bình chọn của bộ phân loại là một chức năng của độ chính xác của nó. Nó có thể được chứng minh rằng với số lượng bộ phân loại yếu ngày càng tăng, sai số huấn luyện của tổ hợp của phân lớp bootsting giảm

**8.16**

Phác thảo các phương pháp để giải quyết vấn đề mất cân bằng lớp học. Giả sử ngân hàng muốn phát triển một bộ phân loại bảo vệ chống lại các giao dịch thẻ tín dụng gian lận. Minh họa cách bạn có thể tạo ra một bộ phân loại chất lượng dựa trên một tập hợp lớn các ví dụ không liên quan và một tập hợp rất nhỏ các trường hợp gian lận.

Trả lời:

Có tổng cộng bốn cách tiếp cận chung để cải thiện độ chính xác phân loại của dữ liệu mất cân bằng phân loại như dữ liệu giao dịch thẻ tín dụng gian lận. Cách đầu tiên là lấy mẫu quá mức, hoạt động bằng cách lấy mẫu lại các bộ giá trị tích cực để kết quả tập huấn luyện chứa một số lượng các bộ giá trị âm và dương bằng nhau. Vì vậy, đối với dữ liệu trong câu hỏi, chúng tôi có thể sao chép bộ mã của các trường hợp gian lận để tạo thành một tập hợp đào tạo có kích thước tương đương với các ví dụ không gian lận.

Cái thứ hai là lấy mẫu dưới, hoạt động bằng cách giảm số lượng bộ giá trị âm. Nó loại bỏ ngẫu nhiên các bộ giá trị khỏi lớp đa số (âm) cho đến khi có số lượng bộ giá trị âm và dương bằng nhau. Vì vậy, đối với vấn đề gian lận trong giao dịch thẻ tín dụng, chúng ta cần loại bỏ ngẫu nhiên các ví dụ không gian lận cho đến khi có quy mô tương đương với các trường hợp gian lận.

Cách thứ ba là di chuyển ngưỡng, áp dụng cho các bộ phân loại, được cung cấp một bộ giá trị đầu vào, trả về giá trị đầu ra liên tục. Thay vì thao tác các bộ giá trị đào tạo, phương thức này trả về quyết định phân loại dựa trên các giá trị đầu ra. Bằng cách điều chỉnh ngưỡng của giá trị đầu ra quyết định dự đoán chính xác của đầu vào, các bộ giá trị lớp hiếm gặp như các trường hợp gian lận trong câu hỏi dễ dàng hơn

phân loại. Và do đó, ít có khả năng xảy ra sai sót âm tính giả tốn kém hơn. Nói cách khác, bộ phân loại sẽ khó phân loại sai các trường hợp gian lận.

Phương pháp cuối cùng là phương pháp Ensemble, được áp dụng cho các thuật toán như tăng cường và rừng ngẫu nhiên. Các bộ phân loại riêng lẻ trong các thuật toán này có thể bao gồm các phiên bản của các phương pháp được mô tả ở trên, chẳng hạn như lấy mẫu quá mức và di chuyển ngưỡng

1. Bài 9.4 trang 439 (chỉ áp dụng cho các bộ phân lớp đã học)

So sánh những ưu điểm và nhược điểm của phân loại eager (ví dụ: cây quyết định, Bayesian, mạng nơ-ron) so với phân loại lười biếng (ví dụ: hàng xóm gần nhất k, dựa trên trường hợp lý luận).

Trả lời:

Eager classification: phân loại nhanh hơn so với phân loại lười biếng vì nó xây dựng một mô hình tổng quát hóa trước khi nhận bất kỳ bộ giá trị mới nào để phân loại. Trọng số có thể được gán cho các thuộc tính, điều này có thể cải thiện độ chính xác của phân loại. Nhược điểm của phân loại háo hức là nó phải cam kết một giả thuyết duy nhất bao trùm toàn bộ không gian cá thể, điều này có thể làm giảm phân loại và cần nhiều thời gian hơn để đào tạo

Lazy classification: sử dụng không gian giả thuyết phong phú hơn, có thể cải thiện độ chính xác của phân loại. Nó đòi hỏi ít thời gian hơn cho việc đào tạo so với phân loại háo hức. Một nhược điểm của phân loại lười biếng là tất cả các bộ huấn luyện cần được lưu trữ, điều này dẫn đến chi phí lưu trữ đắt đỏ và đòi hỏi các kỹ thuật lập chỉ mục hiệu quả. Một nhược điểm khác là phân loại chậm hơn vì các bộ phân loại không được xây dựng cho đến khi các bộ dữ liệu mới cần được phân loại. Hơn nữa, tất cả các thuộc tính đều có trọng số như nhau, điều này có thể làm giảm độ chính xác của phân loại. (Các vấn đề có thể phát sinh do các thuộc tính không liên quan trong dữ liệu.)

1. Bài 9.5 trang 439

Viết một thuật toán cho phân loại k-nearest-neighbor cho k, số lượng gần nhất của hàng xóm, và n, số lượng thuộc tính mô tả mỗi tuple.

1. Bài 10.2 trang 492

Giả sử nhiệm vụ khai thác dữ liệu là cluster points (with (x, y) đại diện cho vị trí) thành ba cụm, trong đó các điểm

A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9).

Chức năng khoảng cách là khoảng cách Euclide. Giả sử ban đầu chúng tôi gán A1, B1 và C1 là trung tâm của mỗi cụm, tương ứng. Sử dụng thuật toán k-means để chỉ hiển thị

(a) Ba cụm tập trung sau vòng thực hiện đầu tiên.

(b) Ba cụm cuối cùng.

1. Bài 10.14

Phân cụm được công nhận là một nhiệm vụ khai thác dữ liệu quan trọng với các ứng dụng rộng lớn. cho một ví dụ ứng dụng cho từng trường hợp sau:

(a) Một ứng dụng sử dụng phân cụm làm chức năng khai thác dữ liệu chính.

(b) Một ứng dụng sử dụng phân cụm như một công cụ xử lý sơ bộ để chuẩn bị dữ liệu cho các nhiệm vụ khai thác dữ liệu khác.

# Khai phá các dạng mẫu trong dữ liệu

1. Bài 6.6 trang 273

Một cơ sở dữ liệu có năm giao dịch. Hãy min\_sup = 60% và min\_conf = 80%.

Text

Description automatically generated

(a) Tìm tất cả các mặt hàng thường xuyên sử dụng Apriori và FP-growth, tương ứng. so sánh hiệu quả của hai quá trình khai thác.

(b) Liệt kê tất cả các quy tắc liên kết mạnh mẽ (với hỗ trợ s và độ tin cậy

c) Thoả mãn quy tắc sau đây, trong đó X là một biến đại diện cho khách hàng và itemi biểu thị các biến đại diện cho các mục (ví dụ: "A", "B,"):

∀x ∈ transaction, buys(X,item1) ∧ buys(X,item2) ⇒ buys(X,item3) [s,c]

1. Bài 6.14 trang 276

Bảng dự phòng sau đây tóm tắt dữ liệu giao dịch siêu thị, trong đó

hot dogs đề cập đến các giao dịch có chứa hot dogs, hot dogs đề cập đến các giao dịch không chứa hot dogs, hamburgers đề cập đến các giao dịch có chứa hamburgers, and hamburgers đề cập đến các giao dịch không chứa hamburge

Text

Description automatically generated

Table

Description automatically generated

(a) Suppose that the association rule “hot dogs ⇒ hamburgers” is mined. Given a

ngưỡng hỗ trợ tối thiểu là 25% và ngưỡng tin cậy tối thiểu là

50%, quy tắc liên kết này có mạnh mẽ không

(b) Based on the given data là việc mua hotdog không phụ thuộc vào việc mua hamburgers? Nếu không, loại mối quan hệ tương quan nào tồn tại giữa cả hai?

(c) So sánh việc sử dụng tất cả sự tự tin, độ tin cậy tối đa, Kulczynski và cosin các biện pháp với lift và mối tương quan trên dữ liệu nhất định

# Thảo luận: Các ứng dụng của Data mining

Thực hành sử dụng phần mềm công cụ Data mining Weka.

**Tài liệu tham khảo:**

[1] J. Han, M. Kamber, J. Pei (2011). Data Mining: Concepts and Techniques. Nhà xuất bản Morgan Kaufmann.