Chương 2

Dữ liệu, phép đo và tiền xử lý dữ liệu

Để tiến hành khai phá dữ liệu thành công, điều quan trọng đầu tiên là làm quen với dữ liệu của bạn. Bạn có thể muốn biết những điều sau: Các loại *thuộc tính* hoặc trường nào tạo nên dữ liệu của bạn? Mỗi thuộc tính có những giá trị như thế nào? Các giá trị này được phân bố ra sao? Làm thế nào để đo lường mức độ tương đồng giữa một số đối tượng dữ liệu với các đối tượng khác? Việc nắm bắt những thông tin này về dữ liệu sẽ giúp ích cho quá trình phân tích sau này. Hơn nữa, dữ liệu thực tế thường chứa nhiều nhiễu, có dung lượng lớn (thường là vài gigabyte hoặc hơn nữa), và có thể xuất phát từ nhiều nguồn không đồng nhất. Làm thế nào để đo lường chất lượng của dữ liệu? Làm sao để làm sạch và tích hợp dữ liệu từ nhiều nguồn khác nhau? Làm thế nào để chuẩn hóa, nén hoặc biến đổi dữ liệu? Làm thế nào để giảm chiều dữ liệu nhằm hỗ trợ phân tích sau này? Đây chính là những nhiệm vụ của chương này.

Chúng ta bắt đầu ở Mục 2.1 bằng cách nghiên cứu các loại thuộc tính khác nhau. Bao gồm: thuộc tính danh định, thuộc tính nhị phân, thuộc tính thứ tự và thuộc tính số. Các *mô tả thống kê* cơ bản có thể được sử dụng để tìm hiểu thêm về các giá trị của từng thuộc tính, như được mô tả trong Mục 2.2. Ví dụ, với thuộc tính *nhiệt độ*, ta có thể xác định *giá trị trung bình*, *trung vị* và *mode* (giá trị xuất hiện nhiều nhất). Đây là các *thước đo của xu hướng trung tâm*, giúp ta có cái nhìn về "điểm giữa" hay trung tâm của một phân phối. Biết được các thống kê cơ bản của từng thuộc tính sẽ giúp cho việc điền các giá trị bị thiếu, làm mượt các giá trị nhiễu, và phát hiện các điểm ngoại lệ trong quá trình tiền xử lý dữ liệu. Kiến thức về các thuộc tính và giá trị của chúng cũng có thể giúp khắc phục những bất nhất (tính không nhất quán) phát sinh trong quá trình tích hợp dữ liệu. Việc vẽ biểu đồ các thước đo xu hướng trung tâm sẽ cho ta biết liệu dữ liệu đối xứng hay lệch. Các biểu đồ như biểu đồ phân vị, biểu đồ tần xuất, và biểu đồ chấm cũng là những cách trực quan hóa các thống kê cơ bản của dữ liệu. Tất cả những công cụ này đều có ích trong quá trình tiền xử lý và cung cấp cái nhìn sâu sắc cho quá trình khai phá.

Chúng ta cũng có thể muốn xem xét mức độ *tương đồng* (hoặc *không tương đồng*) giữa các đối tượng dữ liệu. Ví dụ, giả sử chúng ta có một cơ sở dữ liệu trong đó các đối tượng dữ liệu là bệnh nhân, được mô tả bởi các triệu chứng của họ. Ta có thể muốn tìm mức độ tương đồng hoặc không tương đồng giữa các bệnh nhân riêng lẻ. Thông tin như vậy cho phép ta tìm ra các cụm bệnh nhân tương đồng trong tập dữ liệu. Mức độ tương đồng (hoặc không tương đồng) giữa các đối tượng cũng có thể được sử dụng để phát hiện các điểm ngoại lệ trong dữ liệu, hoặc để thực hiện phân loại theo phương pháp $k-$láng giềng gần nhất. Có rất nhiều thước đo để đánh giá mức độ tương đồng và không tương đồng. Nói chung, những thước đo này được gọi là *thước đo gần gũi* (proximity measure). Hãy nghĩ về mức độ gần gũi của hai đối tượng như là một hàm số của *khoảng cách* giữa các giá trị thuộc tính của chúng, mặc dù mức độ gần gũi cũng có thể được tính toán dựa trên xác suất thay vì khoảng cách thực tế. Các thước đo mức độ gần gũi của dữ liệu được mô tả trong Mục 2.3.

Cuối cùng, chúng ta sẽ thảo luận về tiền xử lý dữ liệu, giải quyết những thách thức thực tế hiện nay: các tập dữ liệu rất dễ bị nhiễu, có giá trị bị thiếu, và không nhất quán do kích thước khổng lồ của chúng và nguồn gốc có thể đến từ nhiều nguồn không đồng nhất. Dữ liệu kém chất lượng sẽ dẫn đến kết quả khai phá kém chất lượng. Rất nhiều nỗ lực cần được đầu tư vào việc tiền xử lý dữ liệu nhằm nâng cao chất lượng dữ liệu cho việc khai phá hiệu quả. Mục 2.4 tập trung vào *làm sạch* và *tích hợp dữ liệu*. Phần làm sạch dữ liệu nhằm loại bỏ nhiễu và sửa các tính

không nhất quán trong dữ liệu, trong khi phần tích hợp dữ liệu nhằm hợp nhất dữ liệu từ nhiều nguồn thành một kho dữ liệu thống nhất, chẳng hạn như kho dữ liệu. Mục 2.5 bàn về *biến đổi dữ liệu*, quá trình biến đổi hoặc hợp nhất dữ liệu thành các dạng phù hợp để khai phá. Nghĩ cách khác, bước này có thể làm cho quá trình khai phá sau này trở nên hiệu quả hơn, và các mẫu tìm được trở nên dễ hiểu hơn. Nhiều chiến lược biến đổi dữ liệu đã được phát triển. Ví dụ, *chuẩn hóa dữ liệu* giúp co giãn các giá trị thuộc tính được đưa về một khoảng giá trị nhỏ hơn, như từ 0.0 đến 1.0; *rời rạc hóa dữ liệu* thay thế các giá trị thô của một thuộc tính số bằng các nhãn khoảng hay nhãn khái niệm; và các kỹ thuật giảm dữ liệu (ví dụ: *nén* và *lấy mẫu*) biến đổi dữ liệu đầu vào thành một biểu diễn rút gọn, có thể cải thiện độ chính xác và hiệu quả của các thuật toán khai phá liên quan đến đo khoảng cách. Cuối cùng, Mục 2.6 bàn về *giảm chiều dữ liệu*, quá trình giảm số lượng biến ngẫu nhiên hay thuộc tính cần xem xét. Xin lưu ý rằng các kỹ thuật tiền xử lý dữ liệu khác nhau không loại trừ lẫn nhau; chúng có thể hoạt động cùng nhau. Ví dụ, làm sạch dữ liệu có thể bao gồm các bước biến đổi để sửa chữa dữ liệu sai, chẳng hạn như chuyển đổi tất cả các mục nhập của một trường *ngày tháng* về cùng một định dạng chung.

2.1 Kiểu dữ liệu

Các tập dữ liệu được tạo thành từ các đối tượng dữ liệu. Một **đối tượng dữ liệu** đại diện cho một thực thể — trong cơ sở dữ liệu bán hàng, các đối tượng có thể là khách hàng, mặt hàng của cửa hàng và giao dịch; trong cơ sở dữ liệu y tế, các đối tượng có thể là bệnh nhân; trong cơ sở dữ liệu của trường đại học, các đối tượng có thể là sinh viên, giáo sư và khóa học. Các đối tượng dữ liệu thường được mô tả bởi các thuộc tính. Các đối tượng dữ liệu cũng có thể được gọi là *mẫu*, *ví dụ*, *điểm dữ liệu*, hoặc *đối tượng*. Nếu các đối tượng dữ liệu được lưu trữ trong cơ sở dữ liệu, chúng được gọi là các *bộ dữ liệu*. Điều này có nghĩa là các hàng của cơ sở dữ liệu tương ứng với các đối tượng dữ liệu, và các cột tương ứng với các thuộc tính. Trong mục này, chúng ta định nghĩa các thuộc tính và xem xét các loại thuộc tính khác nhau.

Thuộc tính là gì? Một **thuộc tính** là một trường dữ liệu, đại diện cho một đặc điểm hoặc tính chất của một đối tượng dữ liệu. Các danh từ như *thuộc tính*, *chiều*, *đặc trưng* và *biến* thường được sử dụng thay thế cho nhau trong các tài liệu nghiên cứu. Thuật ngữ *chiều* thường được sử dụng trong kho dữ liệu. Trong tài liệu về học máy, thường dùng thuật ngữ *đặc trưng*, trong khi các nhà thống kê ưa thích từ

biến. Các chuyên gia về khai phá dữ liệu và cơ sở dữ liệu thường sử dụng từ *thuộc tính*, và chúng tôi cũng sử dụng thuật ngữ này. Các thuộc tính mô tả một đối tượng khách hàng có thể bao gồm, ví dụ, *customer_ID* (mã định danh khách hàng), *name* (tên) và *address* (địa chỉ). Các giá trị quan sát được cho một thuộc tính nhất định được gọi là *các quan sát*. Một tập các thuộc tính được sử dụng để mô tả một đối tượng nhất định được gọi là *véctơ thuộc tính* (hoặc *véctơ đặc trưng*). Phân phối của dữ liệu liên quan đến một thuộc tính (hoặc biến) được gọi là phân phối đơn biến. Một phân phối hai biến liên quan đến hai thuộc tính, và cứ thế.

Kiểu của một thuộc tính được xác định bởi tập các giá trị có thể có của nó – danh định, nhị phân, thứ tự hoặc số – mà thuộc tính đó có thể nhận được. Trong các phần tiếp theo, chúng tôi sẽ giới thiệu từng loại.

2.1.1 Thuộc tính danh định

Danh định có nghĩa là "liên quan đến tên gọi". Các giá trị của một **thuộc tính danh định** là các ký hiệu hoặc *tên gọi của sự vật*. Mỗi giá trị đại diện cho một loại danh mục, mã số hoặc trạng thái nào đó, do đó, các thuộc tính danh định còn được gọi là **thuộc tính phân loại**. Các giá trị này không mang ý nghĩa thứ tự. Trong khoa học máy tính, các giá trị này còn được gọi là các *liệt kê*.

Ví dụ 6 (Thuộc tính danh định)**.** Giả sử rằng *hair_color* (màu tóc) và *marital_status* (tình trạng hôn nhân) là hai thuộc tính mô tả đối tượng *người*. Các giá trị khả dĩ cho *hair_color* ở đây là: *đen*, *nâu*, *vàng*, *đỏ*, *hồng*, *xám* và *trắng*. Thuộc tính *marital_status* có thể nhận các giá trị: *độc thân*, *đã kết hôn*, *ly hôn* và *góa bụa*. Cả *hair_color* và *marital_status* đều là các thuộc tính danh định. Một ví dụ khác của thuộc tính danh định là *occupation* (nghề nghiệp), với các giá trị như *giáo viên*, *nha sĩ*, *lập trình viên*, *nông dân*, v.v.

Mặc dù chúng ta đã nói rằng các giá trị của thuộc tính danh định là các ký hiệu hoặc "tên gọi của sự vật", nhưng có thể biểu diễn các ký hiệu hay "tên" đó bằng các số. Ví dụ, với *hair_color*, chúng ta có thể gán mã 0 cho màu đen, 1 cho màu nâu, v.v. Một ví dụ khác là *customer_ID*, với các giá trị đều là số. Tuy nhiên, trong những trường hợp như vậy, các con số không được dùng theo cách định lượng. Tức là, các phép toán số học trên các giá trị của thuộc tính danh định không có ý nghĩa. Ví dụ, việc trừ một số mã định danh khách hàng cho một số mã định danh khác không có ý nghĩa, không như việc trừ tuổi này cho tuổi khác (trong đó *tuổi* là

thuộc tính số). Mặc dù một thuộc tính danh định có thể có các giá trị là số nguyên, nó không được coi là thuộc tính số bởi vì các số nguyên đó không nhằm mục đích sử dụng định lượng. Chúng ta sẽ nói thêm về các thuộc tính số ở Mục 2.1.4.

Vì các giá trị của thuộc tính danh định không có thứ tự có ý nghĩa và không định lượng, nên việc tìm giá trị trung bình hay trung vị cho một thuộc tính như vậy, dựa trên một tập hợp các đối tượng, là không có ý nghĩa. Tuy nhiên, điều có thể quan tâm là giá trị xuất hiện nhiều nhất của thuộc tính đó. Giá trị này, được gọi là *mode*, là một trong những thước đo của xu hướng trung tâm. Bạn sẽ tìm hiểu thêm về các thước đo xu hướng trung tâm ở Mục 2.2.

2.1.2 Thuộc tính nhị phân

Một **thuộc tính nhị phân** là một thuộc tính danh định chỉ có hai danh mục hoặc trạng thái: 0 hoặc 1, trong đó 0 thường biểu thị rằng thuộc tính không có, còn 1 biểu thị rằng thuộc tính có mặt. Các thuộc tính nhị phân còn được gọi là **Boole** nếu hai trạng thái tương ứng với *đúng* và *sai*.

Ví dụ 7 (Thuộc tính nhị phân)**.** Giả sử thuộc tính *smoker* mô tả một đối tượng *bệnh nhân*, giá trị 1 cho biết bệnh nhân có hút thuốc, trong khi 0 cho biết bệnh nhân không hút thuốc. Tương tự, giả sử bệnh nhân được thực hiện một xét nghiệm y tế với hai kết quả có thể xảy ra. Thuộc tính *medical_test* là thuộc tính nhị phân, trong đó giá trị 1 có nghĩa là kết quả xét nghiệm của bệnh nhân dương tính, còn 0 có nghĩa là kết quả âm tính.

Một thuộc tính nhị phân được gọi là **đối xứng** nếu cả hai trạng thái của nó có giá trị và tầm quan trọng như nhau; tức là, không có sự ưu tiên cho trạng thái nào được mã hóa là 0 hay 1. Một ví dụ như thuộc tính *gender* (giới tính) với các trạng thái *nam* và *nữ*.

Ngược lại, một thuộc tính nhị phân được gọi là **bất đối xứng** nếu các kết quả của hai trạng thái không quan trọng như nhau, chẳng hạn như kết quả *dương tính* và *âm tính* của xét nghiệm HIV. Theo quy ước, chúng ta mã hóa kết quả quan trọng hơn, thường là kết quả hiếm gặp, bằng số 1 (ví dụ: *HIV dương tính*) và kết quả còn lại bằng số 0 (ví dụ: *HIV âm tính*).

Việc tính toán mức độ tương đồng giữa các đối tượng có liên quan đến các thuộc tính nhị phân đối xứng và bất đối xứng sẽ được bàn luận trong một phần sau của chương này.

2.1.3 Thuộc tính thứ tự

Một **thuộc tính thứ tự** là thuộc tính có các giá trị có thứ tự hoặc *xếp hạng* có ý nghĩa với nhau, nhưng khoảng cách về độ lớn giữa các giá trị liên tiếp không được biết rõ.

Ví dụ 8 (Thuộc tính thứ tự)**.** Giả sử thuộc tính *drink_size* tương ứng với kích cỡ của đồ uống có sẵn tại một nhà hàng thức ăn nhanh. Thuộc tính danh định này có ba giá trị khả dĩ: *nhỏ*, *vừa* và *lớn*. Các giá trị này có một trình tự có ý nghĩa (tương ứng với kích cỡ đồ uống tăng dần); tuy nhiên, ta không thể biết được kích cỡ lớn lớn hơn cỡ vừa *bao nhiêu*. Các ví dụ khác của thuộc tính thứ tự bao gồm điểm số (ví dụ: A+, A, A–, B+, v.v.) và thứ bậc nghề nghiệp *professional_rank*. Thứ bậc nghề nghiệp có thể được liệt kê theo một trình tự liên tiếp: ví dụ, đối với giáo sư, các bậc có thể là *trợ giảng*, *phó giáo sư* và *giáo sư*; đối với quân đội, các bậc có thể là *lính mới*, *lính hạng hai* (binh nhì), *lính hạng nhất* (binh nhất), *chuyên viên*, *hạ sĩ*, *sĩ quan*,...
Các thuộc tính thứ tự hữu ích trong việc ghi nhận đánh giá chủ quan về những phẩm chất không thể đo lường một cách khách quan; do đó, chúng thường được sử dụng trong các cuộc khảo sát để đánh giá. Trong một khảo sát, người tham gia có thể được yêu cầu đánh giá mức độ hài lòng của họ với tư cách là khách hàng. Mức độ hài lòng của khách hàng có thể được phân loại theo các danh mục thứ tự sau: *1: rất không hài lòng*, *2: không hài lòng*, *3: trung tính*, *4: hài lòng*, và *5: rất hài lòng*.

Các thuộc tính thứ tự cũng có thể được thu được thông qua việc rời rạc hóa các đại lượng số bằng cách chia khoảng giá trị thành một số danh mục có thứ tự hữu hạn, như được mô tả ở phần sau về giảm dữ liệu.

Xu hướng trung tâm của một thuộc tính thứ tự có thể được thể hiện qua mode và trung vị (giá trị ở giữa của một chuỗi có thứ tự), nhưng không thể xác định được giá trị trung bình.

Lưu ý rằng các thuộc tính danh định, nhị phân và thứ tự là thuộc tính *định tính*. Nghĩa là, chúng *mô tả* một đặc điểm của một đối tượng mà không cho biết độ lớn hay số lượng cụ thể. Các giá trị của những thuộc tính định tính này thường là các từ ngữ biểu thị các danh mục. Nếu sử dụng số nguyên, chúng đại diện cho các mã máy tính cho các danh mục, chứ không phải là các đại lượng có thể đo lường được (ví dụ: 0 cho kích cỡ đồ uống nhỏ, 1 cho cỡ vừa, và 2 cho cỡ lớn). Trong phần tiếp theo, chúng ta sẽ xem xét các thuộc tính số, cung cấp các phép đo *định lượng* của

một đối tượng.

2.1.4 Thuộc tính số

Thuộc tính số mang tính *định lượng*; nghĩa là, nó là một đại lượng có thể đo lường được, được biểu diễn dưới dạng số nguyên hoặc số thực. Các thuộc tính số có thể được chia thành hai loại: thuộc tính theo *thang đo khoảng* và thuộc tính theo *thang đo tỷ lệ*

2.1.4.1 Thuộc tính theo thang đo khoảng

Các **thuộc tính theo thang đo khoảng** được đo trên một thang đo với các đơn vị có kích thước bằng nhau. Các giá trị của thuộc tính này có thứ tự và có thể mang giá trị dương, 0 hoặc âm. Do đó, ngoài việc cung cấp thứ tự giữa các giá trị, chúng còn cho phép chúng ta so sánh và định lượng sự *sai khác* giữa các giá trị.

Ví dụ 9 (Thuộc tính theo thang đo khoảng)**.** Thuộc tính *nhiệt độ* là một ví dụ về thuộc tính theo thang đo khoảng. Giả sử chúng ta có các giá trị *nhiệt độ* ngoài trời cho một số ngày khác nhau, trong đó mỗi ngày được xem là một đối tượng. Khi sắp xếp các giá trị này, ta thu được thứ tự của các đối tượng theo *nhiệt độ*. Ngoài ra, ta cũng có thể định lượng sự chênh lệch giữa các giá trị. Ví dụ, nhiệt độ 20 °C cao hơn 15 °C năm độ. Một ví dụ khác, ngày tháng trong lịch. Chẳng hạn, khoảng cách giữa năm 2012 và năm 2020 là tám năm.

Tuy nhiên, các thang đo nhiệt độ theo độ Celsius và Fahrenheit không có điểm gốc (zero – point) thực sự. Nghĩa là, 0 °C hoặc 0 ° F không có nghĩa là “không có nhiệt độ”. (Trong thang đo Celsius, đơn vị đo được xác định là 1/100 khoảng chênh lệch giữa nhiệt độ nóng chảy và nhiệt độ sôi của nước ở áp suất khí quyển.) Mặc dù chúng ta có thể tính toán sự *chênh lệch* giữa các giá trị nhiệt độ, nhưng không thể nói một giá trị nhiệt độ là *bội số* của một giá trị khác. Chẳng hạn, ta không thể nói rằng 10 °C ấm gấp đôi 5 °C. Điều này là do không có điểm 0 thực sự, nên không thể so sánh theo tỷ lệ. Tương tự, không có điểm 0 thực sự cho ngày tháng trong lịch. (Năm 0 không tương ứng với sự khởi đầu của thời gian.) Điều này dẫn đến khái niệm thuộc tính theo thang đo tỷ lệ, nơi mà một điểm 0 thực sự tồn tại.

Do thuộc tính theo thang đo khoảng là thuộc tính số, nên ta có thể tính toán giá trị trung bình của chúng, ngoài ra còn có thể tính trung vị và mode như các đại lượng thể hiện xu hướng trung tâm.

2.1.4.2 Thuộc tính theo thang đo tỷ lệ

Thuộc tính theo thang đo tỷ lệ là một thuộc tính số có điểm 0 thực sự. Nghĩa là, nếu một phép đo được đo theo thang tỷ lệ, chúng ta có thể nói một giá trị là bội số (hoặc tỷ lệ) của một giá trị khác. Ngoài ra, các giá trị này có thứ tự, và chúng ta có thể tính toán sự khác biệt giữa các giá trị cũng như các đại lượng trung tâm như trung bình, trung vị và mode.

Ví dụ 10 (Thuộc tính theo thang đo tỷ lệ)**.** Không giống như nhiệt độ theo độ Celsius và Fahrenheit, thang đo nhiệt độ Kelvin (K) có điểm 0 thực sự ($0\,\mathrm{K} = -273.15\,^{\circ}\mathrm{C}$). Đây là điểm mà tại đó mọi chuyển động nhiệt dừng lại theo mô tả cổ điển của nhiệt động lực học. Các ví dụ khác về thuộc tính theo thang đo tỷ lệ bao gồm các thuộc tính *đếm* được, chẳng hạn như *years_of_experience* (số năm kinh nghiệm, ví dụ: đối tượng là nhân viên) và *number_of_words* (số lượng từ trong một tài liệu, ví dụ: đối tượng là tài liệu văn bản). Các ví dụ bổ sung bao gồm các thuộc tính đo lường trọng lượng, chiều cao, tốc độ và các đại lượng tiền tệ (ví dụ: nếu bạn có $100 thì bạn giàu hơn 100 lần so với khi có $1).

2.1.5 Thuộc tính rời rạc và thuộc tính liên tục

Trong phần trình bày của chúng ta, các thuộc tính được tổ chức thành các loại danh định, nhị phân, thứ tự và số. Tuy nhiên, có nhiều cách khác nhau để phân loại thuộc tính, và các loại này không hoàn toàn loại trừ lẫn nhau.

Các thuật toán phân loại phát triển từ lĩnh vực học máy thường coi thuộc tính là *rời rạc* hoặc *liên tục*. Mỗi loại thuộc tính có thể được xử lý theo cách khác nhau. **Thuộc tính rời rạc** có tập hợp giá trị hữu hạn hoặc vô hạn đếm được, có thể hoặc không thể biểu diễn dưới dạng số nguyên. Các thuộc tính như *hair_color* (màu tóc), *smoker* (hút thuốc), *medical_test* (kết quả xét nghiệm y tế), *drink_size* (kích thước đồ uống) đều có số lượng giá trị hữu hạn, vì vậy chúng là thuộc tính rời rạc. Lưu ý rằng thuộc tính rời rạc có thể có giá trị số, chẳng hạn như 0 và 1 đối với thuộc tính nhị phân hoặc các giá trị 0 đến 110 cho thuộc tính tuổi *age*. Một thuộc tính được gọi là *vô hạn đếm được* nếu tập hợp các giá trị có thể có là vô hạn nhưng có thể đặt được tương ứng một–một với tập hợp số tự nhiên. Ví dụ, thuộc tính *customer_ID* (mã khách hàng) là vô hạn đếm được. Số lượng khách hàng có thể tăng đến vô hạn, nhưng trong thực tế, tập hợp giá trị của nó vẫn có thể đếm được (có thể đặt tương ứng một–một với tập số nguyên). Một ví dụ khác là mã bưu điện (zip code).

Thuộc tính liên tục là thuộc tính không rời rạc. Trong tài liệu khoa học, thuật ngữ *thuộc tính số* và *thuộc tính liên tục* thường được sử dụng thay thế cho nhau. (Điều này có thể gây nhầm lẫn, vì theo định nghĩa cổ điển, giá trị liên tục là số thực, trong khi giá trị số có thể là cả số nguyên hoặc số thực). Trên thực tế, các giá trị thực được biểu diễn bằng một số lượng chữ số hữu hạn. Các thuộc tính liên tục thường được biểu diễn dưới dạng biến dấu phảy động (floating – point variables).

2.2 Thống kê dữ liệu

Để tiền xử lý dữ liệu thành công, điều quan trọng là phải có một cái nhìn tổng quan về dữ liệu của bạn. Các mô tả thống kê cơ bản có thể được sử dụng để xác định các đặc tính của dữ liệu và làm nổi bật những giá trị cần được coi là nhiễu hoặc ngoại lệ.

Phần này thảo luận về ba lĩnh vực mô tả thống kê cơ bản. Các *thước đo xu hướng trung tâm* (Mục 2.2.1) đo lường vị trí trung tâm của phân phối dữ liệu. Trực quan mà nói, với một thuộc tính, hầu hết các giá trị của nó rơi vào đâu? Các thước đo chính bao gồm *trung bình*, *trung vị*, *mode* và *trung bình khoảng* (midrange).

Các thước đo *độ phân tán của dữ liệu* (Mục 2.2.2) giúp đánh giá mức độ phân bố của dữ liệu. Các thước đo phổ biến nhất bao gồm *khoảng*, *tứ phân vị* (ví dụ, Q_1 là tứ phân vị thứ nhất, tức là bách phân vị thứ 25) và *khoảng tứ phân vị* (interquartile range, IQR). Các khái niệm như *tóm tắt năm số*, biểu đồ hộp, *phương sai* và *độ lệch chuẩn* cũng được sử dụng để phát hiện ngoại lệ.

Các thước đo mối quan hệ giữa nhiều biến số (Mục 2.2.3) bao gồm *hiệp phương sai* và *hệ số tương quan* đối với dữ liệu số, cũng như *kiểm định tương quan* χ^2 (chi – bình phương) đối với dữ liệu danh định.

Biểu diễn trực quan dữ liệu (Mục 2.2.4) nhằm kiểm tra dữ liệu bằng cách sử dụng các biểu đồ. Các công cụ phổ biến trong phần mềm thống kê hoặc trình bày dữ liệu đồ họa bao gồm biểu đồ cột, biểu đồ bánh, biểu đồ đường. Ngoài ra, các phương pháp hiển thị tóm tắt dữ liệu và phân phối phổ biến khác gồm *biểu đồ tần suất*, *biểu đồ phân tán*, *biểu đồ phân vị*, và *biểu đồ phân vị kép* (biểu đồ quantile – quantile).

2.2.1 Đo lường xu hướng trung tâm

Trong phần này, chúng ta sẽ xem xét các cách khác nhau để đo lường xu hướng trung tâm của dữ liệu. Giả sử có một thuộc tính X, chẳng hạn như *mức lương*, đã

được ghi nhận cho một tập hợp các đối tượng. Gọi $x_1, x_2, \ldots, x_n$ là tập hợp gồm n giá trị *quan sát* được của X. Các giá trị này cũng có thể được coi là tập dữ liệu của X. Nếu vẽ biểu đồ các quan sát về *mức lương*, phần lớn các giá trị sẽ rơi vào đâu? Điều này giúp chúng ta hình dung về xu hướng trung tâm của dữ liệu. Các thước đo xu hướng trung tâm bao gồm trung bình, trung vị, mode, và trung bình khoảng (midrange).

Thước đo phổ biến và hiệu quả nhất để xác định "trung tâm" của một tập dữ liệu số là "trung bình số học". Giả sử tập dữ liệu gồm n giá trị *quan sát* được, như *lương* của một nhóm nhân viên, thì **trung bình** của tập giá trị này là

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{2.1}$$

Trong các hệ thống cơ sở dữ liệu quan hệ, *trung bình* được tính bằng hàm tổng hợp có sẵn như *average* (`avg()` trong SQL).

Ví dụ 11 (Trung bình)**.** Giả sử ta có các giá trị mức *lương* (đơn vị: nghìn đô la) sau, được sắp xếp theo thứ tự tăng dần: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Sử dụng phương trình (2.1), ta có:

$$\overline{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58.$$

Do đó, mức lương trung bình là 58 000 đô la.

Đôi khi, mỗi giá trị x_i trong một tập hợp có thể đi kèm với một trọng số w_i cho $i = 1, \ldots, N$. Các trọng số này phản ánh mức độ quan trọng, tầm quan trọng hoặc tần suất xuất hiện của các giá trị tương ứng. Trong trường hợp này, chúng ta có thể tính được

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}, \tag{2.2}$$

được gọi là **trung bình số học có trọng số** hoặc **trung bình trọng số**.

Mặc dù trung bình là đại lượng hữu ích nhất để mô tả một tập dữ liệu, nó không phải lúc nào cũng là cách tốt nhất để đo lường trung tâm của dữ liệu. Một vấn đề lớn của trung bình là độ nhạy cảm của nó đối với các giá trị cực đoan (ví dụ, các giá trị ngoại lệ). Ngay cả một số lượng nhỏ các giá trị cực đoan cũng có thể làm sai lệch trung bình. Ví dụ, mức lương trung bình của một công ty có thể bị đẩy lên

đáng kể bởi mức lương của một vài quản lý được trả lương cao. Tương tự, điểm trung bình của một lớp học trong một kỳ thi có thể bị kéo xuống khá nhiều bởi một vài điểm rất thấp. Để giảm thiểu tác động của các giá trị cực đoan, chúng ta có thể sử dụng **trung bình cắt** bớt (trimmed mean), tức là trung bình được tính sau khi loại bỏ các giá trị ở đầu và cuối của tập dữ liệu. Ví dụ, ta có thể sắp xếp các giá trị quan sát được của mức lương và loại bỏ 2% giá trị cao nhất và 2% giá trị thấp nhất trước khi tính trung bình. Tuy nhiên, cần tránh cắt bớt quá nhiều (ví dụ, 20% ở cả hai đầu), vì điều này có thể làm mất đi những thông tin quý giá.

Đối với dữ liệu lệch (không đối xứng), một đại lượng đo lường trung tâm tốt hơn là **trung vị**, tức là giá trị ở giữa của một tập hợp các giá trị được sắp xếp theo thứ tự. Trung vị chính là giá trị phân chia tập dữ liệu thành hai nửa, nửa trên và nửa dưới.

Trong xác suất và thống kê, trung vị thường áp dụng cho dữ liệu số; tuy nhiên, khái niệm này cũng có thể mở rộng cho dữ liệu thứ tự. Giả sử tập dữ liệu gồm n giá trị của một thuộc tính X được sắp xếp theo thứ tự tăng dần. Nếu n là số lẻ, thì trung vị là giá trị ở giữa của tập hợp đã sắp xếp. Nếu n là số chẵn, thì trung vị không duy nhất; nó có thể được xác định bằng hai giá trị ở giữa và bất kỳ giá trị nào nằm giữa chúng. Nếu X là một thuộc tính số, theo quy ước, trung vị thường được lấy là trung bình của hai giá trị ở giữa.

Ví dụ 12 (Trung vị)**.** Hãy tìm trung vị của dữ liệu từ Ví dụ 11. Dữ liệu đã được sắp xếp theo thứ tự tăng dần. Vì có tổng cộng 12 quan sát (số chẵn), nên trung vị không phải là một giá trị duy nhất. Trung vị có thể là bất kỳ giá trị nào nằm giữa hai giá trị trung tâm là 52 và 56 (tức là giá trị thứ sáu và thứ bảy trong danh sách). Theo quy ước, ta lấy trung bình của hai giá trị này làm trung vị, tức là $\frac{52+56}{2} = \frac{108}{2} = 54$. Do đó, trung vị là \$54 000.
Giả sử ta chỉ có 11 giá trị đầu tiên trong danh sách. Khi số lượng giá trị là lẻ, trung vị chính là giá trị ở giữa. Đây là giá trị thứ sáu trong danh sách, với giá trị \$52 000.

Việc tính trung vị có thể tốn kém khi tập dữ liệu có số lượng quan sát lớn. Tuy nhiên, đối với các thuộc tính số, ta có thể *ước lượng* trung vị một cách dễ dàng. Giả sử dữ liệu được nhóm thành các khoảng theo giá trị x_i, và số lần xuất hiện (tần suất) của mỗi khoảng được biết. Ví dụ, nhân viên có thể được nhóm theo mức lương hàng năm trong các khoảng như \$10 001 – 20 000, \$20 001 – 50 000, v.v. (Một ví dụ cụ thể có thể thấy trong bảng dữ liệu của 9). Gọi khoảng chứa trung vị là *khoảng trung vị*. Ta có thể xấp xỉ trung vị của toàn bộ tập dữ liệu (ví dụ, trung

vị của mức lương) bằng phép nội suy sử dụng công thức

$$\text{median} \approx L_1 + \left(\frac{\frac{n}{2} - \left(\sum \text{freq}\right)_l}{\text{freq}_{\text{median}}} \right) \times \text{width}, \tag{2.3}$$

trong đó L_1 là cận dưới của khoảng trung vị, n là tổng số giá trị trong toàn bộ tập dữ liệu, $\left(\sum \text{freq}\right)_l$ là tổng tần số của tất cả các khoảng nhỏ hơn (bên trái) khoảng trung vị, $\text{freq}_{\text{median}}$ là tần số của khoảng trung vị, *width* là độ rộng của khoảng trung vị.

Mode cũng là một thước đo xu hướng trung tâm. **Mode** của một tập dữ liệu là giá trị xuất hiện với tần suất cao nhất so với các giá trị khác trong tập. Do đó, mode có thể được xác định cho cả thuộc tính định tính và định lượng. Có thể có nhiều giá trị có tần suất xuất hiện cao nhất, dẫn đến sự tồn tại của nhiều mode. Nếu tập dữ liệu có một, hai, hoặc ba mode, nó tương ứng được gọi là **đơn mode**, **song mốt**, **tam mốt**. Nhìn chung, một tập dữ liệu có từ hai mốt trở lên được gọi là **đa mốt**.

Ví dụ 13 (Mode)**.** Dữ liệu của Ví dụ 11 có tính song mode. Hai mode của tập dữ liệu là \$52 000 và \$70 000.

Đối với dữ liệu số đơn mốt có độ lệch (bất đối xứng) vừa phải, ta có mối quan hệ thực nghiệm sau:

$$\text{mean} - \text{mode} \approx 3 \times (\text{mean} - \text{median}). \tag{2.4}$$

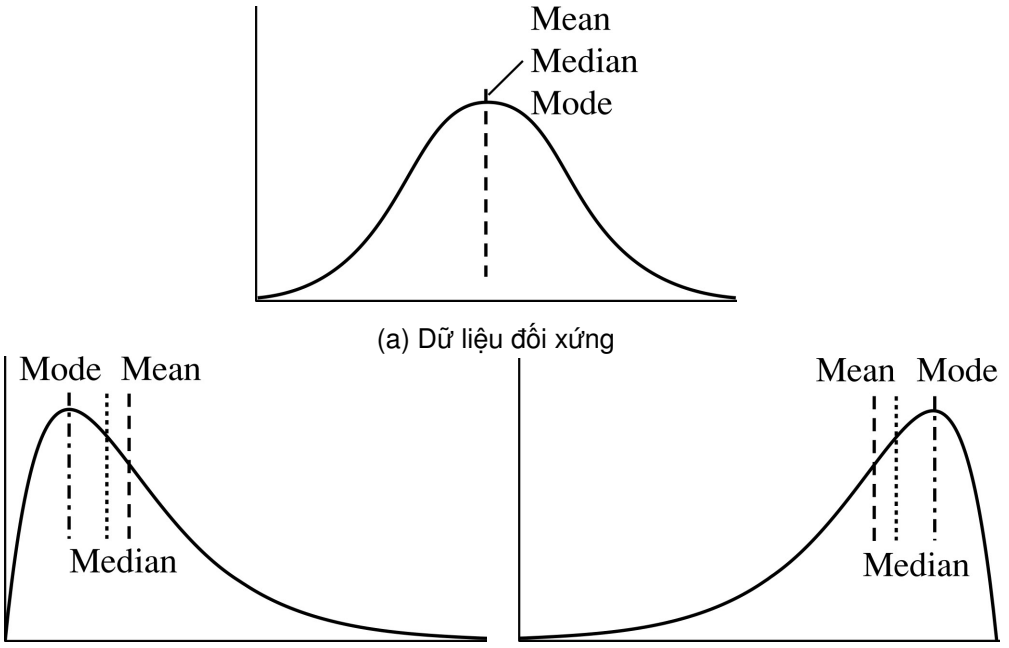
Điều này cho thấy rằng, nếu đã biết giá trị trung bình và trung vị, ta có thể ước tính được mốt của các đường tần số đơn mốt lệch vừa phải.

Trung bình khoảng cũng có thể được sử dụng để đánh giá xu hướng trung tâm của một tập dữ liệu số. Trung bình khoảng là trung bình của giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu. Thước đo này dễ tính toán bằng các hàm tổng hợp trong SQL như `max()` và `min()`.

Ví dụ 14 (Trung bình khoảng)**.** Trung bình khoảng của dữ liệu của Ví dụ 11 là

$$\frac{30\,000 + 110\,000}{2} = \$70\,000.$$

Trong một đường tần suất đơn mode với phân phối dữ liệu hoàn toàn **đối xứng**, trung bình, trung vị và mode đều trùng nhau tại cùng một giá trị trung tâm, như được minh họa trong Hình 2.1a.



(a) Dữ liệu đối xứng

(b) Dữ liệu lệch dương

(c) Dữ liệu lệch âm

Hình 2.1: Trung bình, trung vị và mode của dữ liệu đối xứng so với dữ liệu lệch dương và lệch âm.

Tuy nhiên, dữ liệu trong hầu hết các ứng dụng thực tế không đối xứng. Chúng có thể bị **lệch về phía dương**, khi mode nằm ở giá trị nhỏ hơn trung vị (Hình 2.1b), hoặc bị **lệch về phía âm**, khi mode nằm ở giá trị lớn hơn trung vị (Hình 2.1c).

2.2.2 Đo lường độ phân tán của dữ liệu

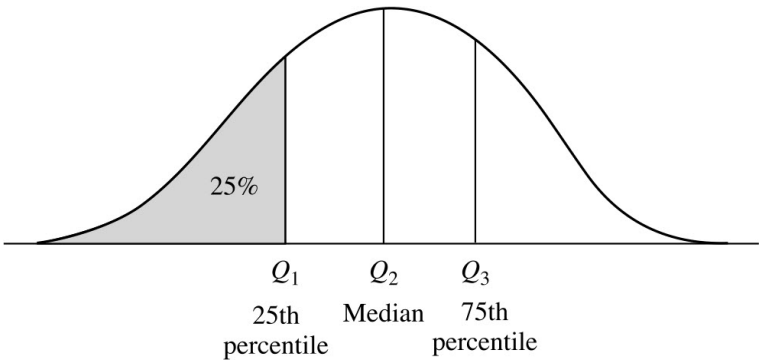
Chúng ta sẽ xem xét các thước đo để đánh giá độ phân tán hay sự trải rộng của dữ liệu số. Các thước đo này bao gồm khoảng, các phân vị, tứ phân vị, bách phân vị và khoảng tứ phân vị (interquartile range – IQR). Tóm tắt năm số, có thể được hiển thị dưới dạng biểu đồ hộp, rất hữu ích trong việc phát hiện các giá trị ngoại lệ. Phương sai và độ lệch chuẩn cũng chỉ ra mức độ phân tán của một phân phối dữ liệu.

2.2.2.1 Khoảng, phân vị, tứ phân vị và khoảng tứ phân vị

Giả sử $x_1, x_2, \ldots, x_n$ là một tập các quan sát của một thuộc tính số X. **Khoảng** của tập dữ liệu là hiệu giữa giá trị lớn nhất (`max()`) và giá trị nhỏ nhất (`min()`).

Giả sử dữ liệu của thuộc tính X được sắp xếp theo thứ tự tăng dần. Hãy tưởng

tượng rằng ta có thể chọn một số điểm dữ liệu nhất định để chia phân phối dữ liệu thành các tập liên tiếp có cỡ bằng nhau, như minh họa trong Hình 2.2. Những điểm dữ liệu này được gọi là *phân vị*. **Phân vị** là các điểm được lấy ở các khoảng cách đều trong một phân phối dữ liệu, chia nó thành các tập liên tiếp có cỡ cơ bản bằng nhau. (Chúng ta nói "cơ bản bằng nhau" vì có thể không tồn tại các giá trị của X chia dữ liệu thành các tập con có cỡ hoàn toàn bằng nhau. Vì mục đích dễ đọc, chúng ta sẽ gọi chúng là bằng nhau.) Phân vị $q-$*phần* thứ k của một phân phối dữ liệu là giá trị x sao cho không quá $\frac{k}{q}$ giá trị của dữ liệu nhỏ hơn x và không quá $\frac{q-k}{q}$ giá trị của dữ liệu lớn hơn x, với k là một số nguyên sao cho $0 < k < q$. Số lượng phân vị $q-$phần là $(q-1)$.



Hình 2.2: Một biểu đồ phân phối dữ liệu của thuộc tính X. Các phân vị được vẽ là các tứ phân vị. Ba tứ phân vị chia phân phối thành bốn tập liên tiếp có cỡ bằng nhau. Tứ phân vị thứ hai tương ứng với trung vị.

Phân vị 2 phần: đây là điểm dữ liệu chia đôi phân phối dữ liệu; nó tương ứng với trung vị. Phân vị 4 phần: đây là ba điểm dữ liệu chia phân phối thành bốn phần bằng nhau; mỗi phần chiếm một phần tư của phân phối. Chúng thường được gọi là **tứ phân vị**. Phân vị 100 phần: thường được gọi là **bách phân vị**; chúng chia phân phối dữ liệu thành 100 tập liên tiếp có cỡ bằng nhau. Trung vị, tứ phân vị và bách phân vị là các hình thức phân vị được sử dụng phổ biến nhất.

Các tứ phân vị cho ta biết về trung tâm, độ trải rộng và hình dạng của một phân phối dữ liệu. **Tứ phân vị thứ nhất** (Q1) là bách phân vị thứ 25; nó cắt bỏ 25% dữ liệu nhỏ nhất. **Tứ phân vị thứ ba** (Q3) là bách phân vị thứ 75; nó cắt bỏ 75% dữ liệu nhỏ nhất (hoặc, nói cách khác, giữ lại 25% dữ liệu lớn nhất). Tứ phân vị thứ hai tương đương với trung vị (bách phân vị thứ 50), cho biết trung tâm của phân phối dữ liệu.

Khoảng tứ phân vị (IQR) là khoảng cách giữa tứ phân vị thứ ba và tứ phân vị thứ nhất, một thước đo đơn giản về mức độ trải rộng của nửa dữ liệu giữa:

$$\text{IQR} = Q_3 - Q_1. \tag{2.5}$$

Ví dụ 15 (Khoảng tứ phân vị)**.** Các tứ phân vị là ba giá trị chia tập dữ liệu đã sắp xếp thành bốn phần bằng nhau. Dữ liệu của Ví dụ 11 gồm 12 quan sát, đã được sắp xếp theo thứ tự tăng dần. Vì danh sách có số phần tử chẵn, nên trung vị của danh sách được tính là trung bình của hai phần tử ở giữa, tức là $\frac{\$52\,000 + \$56\,000}{2} = \$54\,000$. Khi đó, tứ phân vị thứ nhất được tính là trung bình của phần tử thứ 3 và thứ 4, tức là $\frac{\$47\,000 + \$50\,000}{2} = \$48\,500$, trong khi tứ phân vị thứ ba được tính là trung bình của phần tử thứ 9 và thứ 10, tức là $\frac{\$63\,000 + \$70\,000}{2} = \$66\,500$. Như vậy, khoảng tứ phân vị là

$$\text{IQR} = \$66\,500 - \$48\,500 = \$18\,000.$$

2.2.2.2 Tóm tắt năm số, biểu đồ hộp và giá trị ngoại lệ

Không có một thước đo duy nhất nào (ví dụ: *IQR*) đủ hữu ích để mô tả mức độ phân tán của các phân phối lệch. Hãy nhìn vào các phân phối đối xứng và lệch trong Hình 2.1. Trong phân phối đối xứng, trung vị (và các thước đo xu hướng trung tâm khác) chia dữ liệu thành hai nửa có kích thước bằng nhau. Điều này không xảy ra đối với các phân phối lệch. Do đó, thông tin sẽ đầy đủ hơn khi ta cung cấp thêm hai tứ phân vị Q_1 và Q_3 cùng với trung vị. Một quy tắc chung để xác định các giá trị nghi ngờ là ngoại lệ là chọn ra những giá trị nằm cách Q_3 ít nhất $1.5 \times \text{IQR}$ hoặc cách Q_1 ít nhất $1.5 \times \text{IQR}$.

Vì Q_1, trung vị và Q_3 không cung cấp thông tin về các điểm cuối (ví dụ: đuôi của phân phối) của dữ liệu, nên một bản tóm tắt đầy đủ hơn về hình dạng của phân phối có thể được cung cấp bằng cách ghi lại cả giá trị nhỏ nhất và lớn nhất. Đây được gọi là *tóm tắt năm số*. **Tóm tắt năm số** của một phân phối gồm có *giá trị nhỏ nhất*, *tứ phân vị thứ nhất* (Q_1), *trung vị* (Q_2), *tứ phân vị thứ ba* ($Q3$), và *giá trị lớn nhất*.

Biểu đồ hộp là một cách trực quan phổ biến để hiển thị phân phối dữ liệu, dựa trên tóm tắt năm số như sau:

- Thông thường, hai đầu của hộp được đặt tại các tứ phân vị, do đó chiều dài của hộp là khoảng tứ phân vị (IQR).

- Trung vị được đánh dấu bằng một đường nằm bên trong hộp.
- Hai đường kẻ (gọi là "*râu*") mở rộng từ hộp ra đến các giá trị nhỏ nhất và lớn nhất.

Khi xử lý một số lượng quan sát vừa phải, ta nên vẽ các giá trị ngoại lai riêng lẻ. Để làm như vậy trong biểu đồ hộp, các râu sẽ được kéo dài đến các giá trị cực thấp và cực cao chỉ nếu các giá trị này nằm cách các tứ phân vị hơn $1.5 \times$ IQR. Nếu không, các râu sẽ dừng lại tại các giá trị cực đoan gần nhất nằm trong khoảng $1.5 \times$ IQR của các tứ phân vị, và các trường hợp còn lại được vẽ riêng lẻ. Biểu đồ hộp có thể được sử dụng để so sánh nhiều tập dữ liệu tương thích.

Ví dụ 16 (Biểu đồ hộp)**.** Hình 2.3 cho thấy các biểu đồ hộp của dữ liệu đơn giá cho các mặt hàng được bán tại bốn chi nhánh của một cửa hàng trực tuyến trong một khoảng thời gian nhất định. Đối với chi nhánh 1, ta thấy rằng giá trị trung vị của các mặt hàng bán ra là \$80, Q_1 là \$60, và Q_3 là \$100. Lưu ý rằng hai quan sát ngoại lệ của chi nhánh này được vẽ riêng lẻ, vì giá trị của chúng là 175 và 202, vượt quá 1.5 lần IQR, ở đây là 40.

2.2.2.3 Phương sai và độ lệch chuẩn

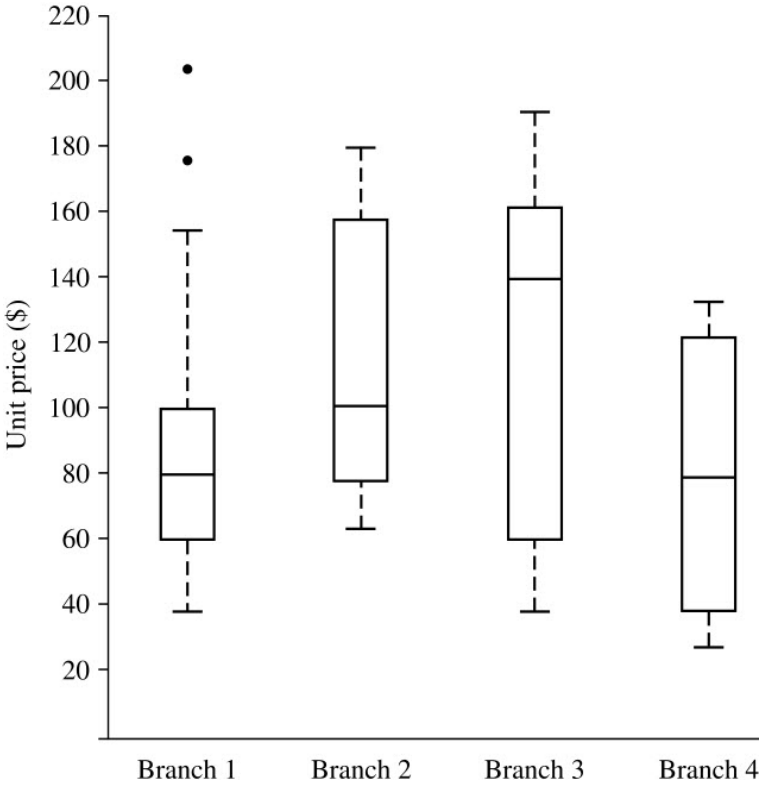
Phương sai và độ lệch chuẩn là các thước đo độ phân tán của dữ liệu. Chúng cho biết mức độ phân tán của một phân bố dữ liệu. Độ lệch chuẩn nhỏ cho thấy các quan sát dữ liệu có xu hướng rất gần với giá trị trung bình, trong khi độ lệch chuẩn lớn chỉ ra rằng dữ liệu được phân bố trên một khoảng giá trị rộng.

Phương sai của n quan sát $x_1, x_2, \ldots, x_n$ (khi n lớn), đối với một thuộc tính số X, là

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2, \tag{2.6}$$

trong đó $\overline{x}$ là giá trị trung bình của các quan sát, được định nghĩa trong công thức (2.1). **Độ lệch chuẩn** σ của các quan sát là căn bậc hai của phương sai σ^2.

Ví dụ 17 (Phương sai và độ lệch chuẩn)**.** Trong Ví dụ 11, ta tìm được $\overline{x} = \$58\,000$ bằng cách sử dụng công thức (2.1) cho giá trị trung bình. Để xác định phương sai và độ lệch chuẩn của dữ liệu từ ví dụ đó, ta đặt $n = 12$ và sử dụng công thức (2.6)



Hình 2.3: Biểu đồ hộp của dữ liệu đơn giá các mặt hàng được bán tại bốn chi nhánh của một cửa hàng trực tuyến trong một khoảng thời gian nhất định.

để tính

$$\begin{aligned}\sigma^2 &= \frac{1}{12}\left(30^2 + 36^2 + 47^2 + \cdots + 110^2\right) - 58^2 \\ &\approx 379.17 \\ \sigma &\approx \sqrt{379.17} \approx 19.47.\end{aligned}$$

Các tính chất cơ bản của độ lệch chuẩn σ với vai trò là thước đo độ phân tán như sau:

- σ đo lường mức độ phân tán quanh giá trị trung bình, và chỉ nên được xem xét khi sử dụng giá trị trung bình làm thước đo trung tâm.
- $\sigma = 0$ chỉ khi không có sự phân tán nào, tức là khi tất cả các quan sát có cùng một giá trị. Ngược lại, nếu có sự khác biệt giữa các quan sát, thì $\sigma > 0$.

Điều quan trọng là một quan sát hiếm khi nằm cách xa trung bình quá nhiều độ lệch chuẩn. Về mặt toán học, sử dụng bất đẳng thức Chebyshev, có thể chứng

minh rằng ít nhất $\left(1 - \frac{1}{k^2}\right) \times 100\%$ các quan sát nằm trong phạm vi không quá k lần độ lệch chuẩn tính từ trung bình. Do đó, độ lệch chuẩn là một chỉ báo tốt về độ phân tán của một tập dữ liệu.

Việc tính toán phương sai và độ lệch chuẩn có khả năng mở rộng tốt đối với các tập dữ liệu lớn.

2.2.3 Phân tích hiệp phương sai và tương quan

2.2.3.1 Hiệp phương sai của dữ liệu số

Trong lý thuyết xác suất và thống kê, tương quan và hiệp phương sai là hai thước đo tương tự nhau dùng để đánh giá mức độ hai thuộc tính thay đổi cùng nhau như thế nào.

Xét hai thuộc tính số X và Y, cùng với một tập gồm n quan sát giá trị thực: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Giá trị trung bình của X và Y lần lượt được gọi là **giá trị kỳ vọng** của X và Y, tức là

$$E(X) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{và} \quad E(Y) = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Hiệp phương sai giữa X và Y được định nghĩa là:

$$\text{Cov}(X, Y) = E[(X - \overline{x})(Y - \overline{y})] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \tag{2.7}$$

Về mặt toán học, cũng có thể chứng minh rằng:

$$\text{Cov}(X, Y) = E(XY) - \overline{x}\,\overline{y}. \tag{2.8}$$

Công thức này giúp đơn giản hóa tính toán.

Nếu hai thuộc tính X và Y có xu hướng thay đổi cùng nhau, nghĩa là khi một giá trị x_i của X lớn hơn trung bình $\overline{x}$, thì giá trị tương ứng y_i của Y cũng có khả năng lớn hơn $\overline{y}$. Khi đó, hiệp phương sai của X và Y là *dương*. Ngược lại, nếu một thuộc tính có giá trị lớn hơn trung bình khi thuộc tính kia nhỏ hơn trung bình, thì hiệp phương sai là *âm*.

Nếu X và Y *độc lập* (không có mối tương quan), thì $E(XY) = E(X) \cdot E(Y)$, do đó $\text{Cov}(X, Y) = E(XY) - \overline{x}\,\overline{y} = E(X) \cdot E(Y) - \overline{x}\,\overline{y} = 0$. Tuy nhiên, điều ngược lại không đúng: một số cặp biến ngẫu nhiên có hiệp phương sai bằng 0 nhưng không độc lập. Chỉ dưới một số giả định bổ sung (ví dụ: dữ liệu tuân theo phân phối chuẩn nhiều chiều) thì hiệp phương sai bằng 0 mới dẫn đến tính độc lập.

Ví dụ 18 (Phân tích hiệp phương sai của các thuộc tính số)**.** Xét Bảng 2.1, trình bày một ví dụ đơn giản về giá cổ phiếu được quan sát tại năm thời điểm khác nhau của công ty *AllElectronics* và công ty *HighTech*, một công ty công nghệ cao. Nếu các cổ phiếu bị ảnh hưởng bởi cùng một xu hướng ngành, thì giá của chúng có xu hướng tăng hoặc giảm cùng nhau hay không?

Thời điểm	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Bảng 2.1: Giá cổ phiếu của *AllElectronics* và *HighTech*.

Lời giải. Tính giá trị kỳ vọng (trung bình) của giá cổ phiếu của *AllElectronics*:

$$E(\text{AllElectronics}) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4.$$

Tính giá trị kỳ vọng cho HighTech:

$$E(\text{HighTech}) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

Sử dụng công thức (2.8), ta tính hiệp phương sai:

$$\begin{aligned} &\text{Cov}(\text{AllElectronics}, \text{HighTech}) \\ &= \frac{6\times 20 + 5\times 10 + 4\times 14 + 3\times 5 + 2\times 5}{5} - 4\times 10.80 \\ &= 50.2 - 43.2 = 7. \end{aligned}$$

Do đó, với hiệp phương sai dương là 7, ta có thể nói rằng giá cổ phiếu của cả hai công ty có xu hướng tăng cùng nhau. □

Phương sai là một trường hợp đặc biệt của hiệp phương sai, khi hai thuộc tính là trùng nhau (tức là hiệp phương sai của một thuộc tính với chính nó).

2.2.3.2 Hệ số tương quan cho dữ liệu số

Đối với các thuộc tính số, chúng ta có thể đánh giá mức độ tương quan giữa hai thuộc tính X và Y bằng cách tính **hệ số tương quan** (còn gọi là hệ số mômen tích Pearson, được đặt theo tên của Karl Pearson – người phát minh ra nó). Đó là

$$r_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n \sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} (x_i y_i) - n \overline{x}\,\overline{y}}{n \sigma_X \sigma_Y}, \tag{2.9}$$

trong đó n là số lượng bộ dữ liệu, x_i và y_i là các giá trị tương ứng của X và Y trong bộ dữ liệu thứ i, $\overline{x}$ và $\overline{y}$ là giá trị trung bình tương ứng của X và Y, σ_X và σ_Y là độ lệch chuẩn tương ứng của X và Y (như được định nghĩa ở Mục 2.2.2), $\sum (x_i y_i)$ là tổng của tích các giá trị X và Y của từng bộ dữ liệu. Lưu ý rằng giá trị của $-1 \leq r_{X,Y} \leq 1$. Nếu $r_{X,Y} > 0$, thì X và Y có *tương quan dương*, có nghĩa là khi giá trị của X tăng, giá trị của Y cũng có xu hướng tăng. Giá trị càng cao thì mức độ tương quan càng mạnh (nghĩa là mỗi thuộc tính càng gợi ý cho thuộc tính kia). Một giá trị tương quan cao có thể cho thấy rằng một trong hai thuộc tính có thể bị loại bỏ do thừa thông tin. Nếu $r_{X,Y} = 0$, thì X và Y là *độc lập*, không có tương quan giữa chúng. Nếu $r_{X,Y} < 0$, thì X và Y có *tương quan âm*, nghĩa là khi giá trị của một thuộc tính tăng, giá trị của thuộc tính kia lại giảm, tức là mỗi thuộc tính lại "chống lại" thuộc tính kia. Biểu đồ phân tán cũng có thể được sử dụng để xem mức độ tương quan giữa các thuộc tính (xem Mục 2.2.3). Ví dụ, Hình 2.8 cho thấy các biểu đồ phân tán của dữ liệu có tương quan dương và tương quan âm, trong khi Hình 2.9 hiển thị dữ liệu không có tương quan.

Lưu ý rằng tương quan không đồng nghĩa với nhân quả. Nói cách khác, nếu X và Y có tương quan, điều này không nhất thiết có nghĩa là X gây ra Y hoặc Y gây ra X. Ví dụ, khi phân tích một cơ sở dữ liệu nhân khẩu học, ta có thể phát hiện rằng số lượng bệnh viện và số vụ trộm xe trong một vùng có tương quan. Điều này không có nghĩa là một yếu tố gây ra yếu tố kia; cả hai thực sự có mối liên hệ nhân quả với một thuộc tính thứ ba, đó là *dân số*.

2.2.3.3 Kiểm định tương quan χ^2 cho dữ liệu danh định

Đối với dữ liệu danh định, mối quan hệ tương quan giữa hai thuộc tính X và Y có thể được phát hiện thông qua kiểm định χ^2 (**chi – square**). Giả sử X có r giá trị khác nhau, cụ thể là $x_1, x_2, \ldots, x_r$, và Y có c giá trị khác nhau, cụ thể là $y_1, y_2, \ldots, y_c$. Các bộ dữ liệu được mô tả bởi X và Y có thể được trình bày dưới dạng một **bảng tiếp liên**, trong đó r giá trị của X tạo thành các hàng và c giá trị của Y tạo thành

các cột. Gọi $\left(A_i, B_j\right)$ là biến cố đồng thời mà thuộc tính X nhận giá trị x_i và thuộc tính Y nhận giá trị y_j (tức là, $X = x_i$, $Y = y_j$). Mỗi biến cố đồng thời $\left(A_i, B_j\right)$ có một ô riêng (hoặc vị trí) trong bảng. Giá trị χ^2 (còn được gọi là *thống kê Pearson* χ^2) được tính bằng công thức:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}, \tag{2.10}$$

trong đó o_{ij} là *tần số quan sát* được (số đếm thực tế) của biến cố đồng thời $\left(A_i, B_j\right)$, e_{ij} là *tần số kỳ vọng* của $\left(A_i, B_j\right)$, được tính theo công thức:

$$e_{ij} = \frac{\text{count}\left(X = x_i\right) \times \text{count}\left(B = b_j\right)}{n}, \tag{2.11}$$

với n là số lượng bộ dữ liệu, $\text{count}\left(X = x_i\right)$ là số bộ có giá trị x_i của X, và $\text{count}\left(Y = y_j\right)$ là số bộ có giá trị y_j của Y. Tổng trong công thức (2.10) được tính trên tất cả $r \times c$ ô của bảng. Lưu ý rằng các ô có sự chênh lệch lớn giữa số đếm thực tế và số đếm kỳ vọng sẽ đóng góp nhiều hơn vào giá trị χ^2.

Thống kê χ^2 được dùng để kiểm định giả thuyết rằng X và Y là *độc lập*, tức là không có tương quan giữa chúng. Kiểm định này dựa trên mức ý nghĩa với số bậc tự do là $(r-1) \times (c-1)$. Chúng ta sẽ minh họa cách sử dụng thống kê này trong Ví dụ 19. Nếu giả thuyết độc lập bị bác bỏ, ta có thể nói rằng X và Y có tương quan thống kê.

Ví dụ 19 (Phân tích tương quan của các thuộc tính danh định sử dụng kiểm định χ^2)**.** Giả sử rằng một nhóm gồm 1500 người đã được khảo sát. Giới tính của mỗi người được ghi nhận. Mỗi người cũng được hỏi về loại tài liệu đọc ưa thích của mình là tiểu thuyết (fiction) hay phi tiểu thuyết (non−fiction). Như vậy, chúng ta có hai thuộc tính: *gender* (giới tính) và *preferred_reading* (sở thích đọc). Số lượng quan sát (hoặc số đếm) của mỗi biến cố đồng thời có thể xảy ra được tóm tắt trong bảng tiếp liên được trình bày trong Bảng 2.2, trong đó các số trong ngoặc đơn là các tần số kỳ vọng. Các tần số kỳ vọng được tính dựa trên phân phối dữ liệu của cả hai thuộc tính theo công thức (2.11).

Lời giải. Sử dụng công thức (2.11), ta có thể kiểm tra các tần số kỳ vọng cho từng ô. Ví dụ, tần số kỳ vọng cho ô (nam, fiction) là:

$$e_{11} = \frac{\text{count (nam)} \times \text{count (fiction)}}{n} = \frac{300 \times 450}{1500} = 90,$$

	Nam	**Nữ**	**Tổng**
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Tổng	**300**	**1200**	**1500**

Lưu ý: Liệu giới tính và sở thích đọc có tương quan không?

Bảng 2.2: Dữ liệu bảng tiếp liên 2 $\times$ 2 của Ví dụ 19.

Và tương tự cho các ô khác. Lưu ý rằng, trong hàng bất kỳ, tổng các tần số kỳ vọng phải bằng tổng số quan sát của hàng đó, và tổng các tần số kỳ vọng ở cột bất kỳ cũng phải bằng tổng số quan sát của cột đó.

Sử dụng công thức (2.10) để tính thống kê χ^2, ta có:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Với bảng 2 $\times$ 2 này, số bậc tự do là $(2-1) \times (2-1) = 1$. Với 1 bậc tự do, giá trị χ^2 cần thiết để bác bỏ giả thuyết độc lập ở mức ý nghĩa 0.001 là 10.828 (theo bảng phân vị phải của phân phối χ^2, thường có trong sách giáo trình thống kê). Vì giá trị tính được của chúng ta vượt quá con số này, ta có thể bác bỏ giả thuyết cho rằng giới tính và và sở thích đọc là độc lập và kết luận rằng hai thuộc tính này có tương quan (mạnh) đối với nhóm người được khảo sát. □

2.2.4 Biểu diễn đồ họa của thống kê cơ bản của dữ liệu

Trong mục này, chúng ta sẽ nghiên cứu các biểu đồ hiển thị các mô tả thống kê cơ bản. Những biểu đồ này bao gồm *biểu đồ phân vị*, *biểu đồ phân vị kép* (biểu đồ quantile–quantile), *biểu đồ tần suất*, và *biểu đồ phân tán*. Những biểu đồ này giúp ta kiểm tra trực quan dữ liệu, rất hữu ích cho việc tiền xử lý. Ba biểu đồ đầu tiên hiển thị phân phối đơn biến (tức là dữ liệu của một thuộc tính), trong khi biểu đồ phân tán hiển thị phân phối hai biến (tức là liên quan đến hai thuộc tính).

2.2.4.1 Biểu đồ phân vị

Biểu đồ phân vị là một cách đơn giản và hiệu quả để có cái nhìn ban đầu về phân phối dữ liệu đơn biến. Đầu tiên, nó hiển thị tất cả dữ liệu của thuộc tính đã cho (cho phép người dùng đánh giá cả hành vi tổng thể và những trường hợp bất

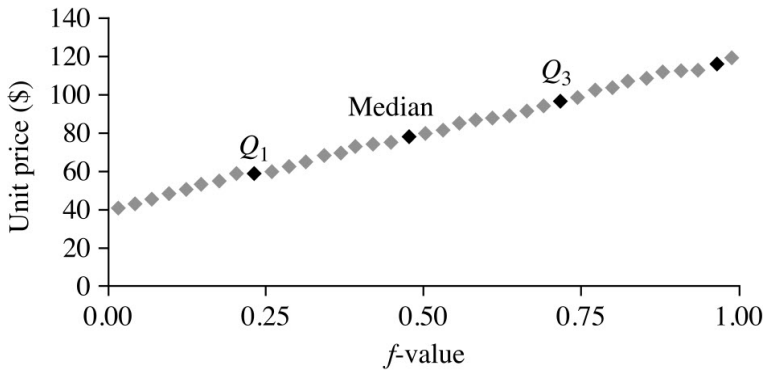
thường). Thứ hai, nó vẽ thông tin phân vị (xem Mục 2.2.2). Giả sử x_i, với $i = 1, \ldots, n$, là các dữ liệu được sắp xếp theo thứ tự tăng dần, sao cho x_1 là quan sát nhỏ nhất và x_n là quan sát lớn nhất cho thuộc tính thứ tự hoặc số X. Mỗi quan sát x_i được ghép với một giá trị phần trăm f_i, cho biết có khoảng $f_i \times 100\%$ dữ liệu nằm bên dưới giá trị x_i. Ta nói "khoảng" vì có thể không tồn tại một giá trị nào mà chính xác có một phần f_i của dữ liệu nằm dưới x_i. Lưu ý rằng, phân vị 0.25 tương ứng với tứ phân vị thứ nhất Q_1, phân vị 0.50 chính là trung vị, và phân vị 0.75 tương ứng với Q_3.

Giả sử ta tính:

$$f_i = \frac{i - 0.5}{n}. \tag{2.12}$$

Những số này tăng dần với bước nhảy bằng $\frac{1}{n}$, dao động từ $\frac{1}{2n}$ (hơi lớn hơn 0) đến $1 - \frac{1}{2n}$ (hơi nhỏ hơn 1). Trên biểu đồ phân vị, x_i được vẽ theo f_i, tức là x_i được vẽ theo trục tung, còn f_i trên trục hoành, hoặc, đôi khi, đảo lại. Điều này cho phép ta so sánh các phân phối khác nhau dựa trên các phân vị của chúng. Ví dụ, dựa vào các biểu đồ phân vị của dữ liệu bán hàng của hai khoảng thời gian khác nhau, ta có thể so sánh nhanh chóng Q_1, trung vị, Q_3 và các giá trị f_i khác.

Ví dụ 20 (Biểu đồ phân vị)**.** Hình 2.4 cho thấy một biểu đồ phân vị của dữ liệu *đơn giá* được trình bày trong Bảng 2.3.



Hình 2.4: Một biểu đồ phân vị cho dữ liệu đơn giá của Bảng 2.3.

2.2.4.2 Biểu đồ phân vị kép

Biểu đồ phân vị kép, hay **biểu đồ quantile – quantile**, **biểu đồ q – q**, vẽ các phân vị của một phân phối đơn biến so với các phân vị tương ứng của một phân

Đơn giá ($)	Số lượng mặt hàng bán ra
40	275
43	300
47	250
...	...
...	...
74	360
75	515
78	540
...	...
...	...
115	320
117	270
120	350

Bảng 2.3: Một tập hợp dữ liệu đơn giá cho các mặt hàng được bán tại một chi nhánh của cửa hàng trực tuyến.

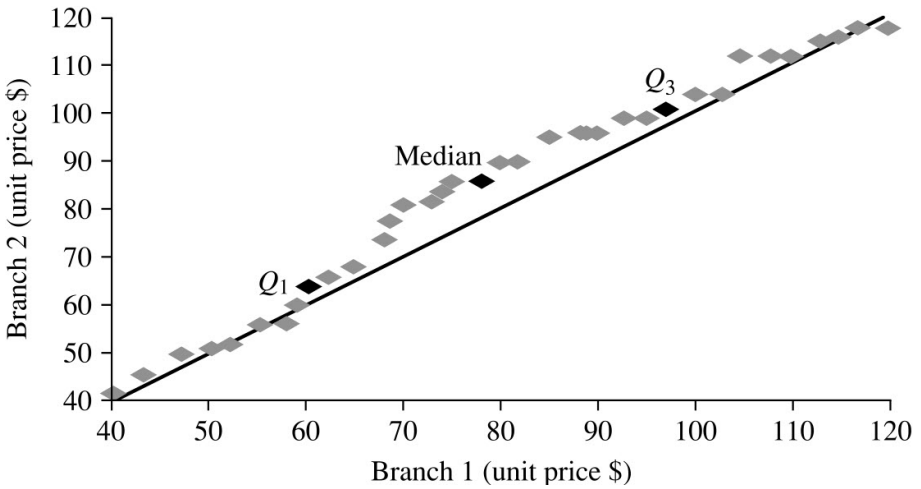
phối khác. Đây là một công cụ trực quan mạnh mẽ vì nó cho phép người dùng xem liệu có sự dịch chuyển nào khi chuyển từ một phân phối sang phân phối khác hay không.

Giả sử ta có hai tập hợp các quan sát cho thuộc tính, hay biến, *đơn giá*, lấy từ hai chi nhánh khác nhau. Gọi $x_1, \dots, x_n$ là dữ liệu từ chi nhánh thứ nhất, và $y_1, \dots, y_m$ là dữ liệu từ chi nhánh thứ hai, với mỗi tập dữ liệu được sắp xếp theo thứ tự tăng dần. Nếu $m = n$ (tức là số lượng điểm trong mỗi tập là bằng nhau), ta chỉ cần vẽ y_i theo x_i, trong đó y_i và x_i đều là phân vị $\frac{i-0.5}{n}$ của tập dữ liệu tương ứng. Nếu $m < n$ (tức là chi nhánh thứ hai có ít quan sát hơn chi nhánh thứ nhất), thì biểu đồ phân vị kép chỉ có m điểm. Ở đây, y_i là phân vị $\frac{i-0.5}{m}$ của dữ liệu Y, và được vẽ đối với phân vị $\frac{i-0.5}{m}$ của dữ liệu X. Quá trình tính toán này thường yêu cầu nội suy.

Ví dụ 21 (Biểu đồ phân vị kép)**.** Hình 2.5 cho thấy một biểu đồ phân vị kép cho dữ liệu *đơn giá* của các mặt hàng được bán tại hai chi nhánh của cửa hàng trực tuyến trong một khoảng thời gian nhất định. Mỗi điểm trên biểu đồ tương ứng với cùng một phân vị của mỗi tập dữ liệu, cho biết đơn giá của các mặt hàng bán ra

ở chi nhánh 1 so với chi nhánh 2 tại phân vị đó. (Để hỗ trợ việc so sánh, đường thẳng trên biểu đồ biểu diễn trường hợp mà, đối với mỗi phân vị, đơn giá ở cả hai chi nhánh là giống nhau. Các điểm tối hơn tương ứng với dữ liệu của Q_1 (tứ phân vị thứ nhất), trung vị và Q_3 (tứ phân vị thứ ba) theo thứ tự.)

Ví dụ, chúng ta nhận thấy ở Q_1, đơn giá của mặt hàng bán tại chi nhánh 1 hơi thấp hơn so với chi nhánh 2. Nói cách khác, 25% số mặt hàng bán ra ở chi nhánh 1 có đơn giá nhỏ hơn hoặc bằng \$60, trong khi 25% số mặt hàng bán ra ở chi nhánh 2 có đơn giá nhỏ hơn hoặc bằng \$64. Ở bách phân vị thứ 50 (được đánh dấu bởi trung vị, cũng chính là Q_2), ta thấy rằng 50% số mặt hàng bán ra ở chi nhánh 1 có đơn giá nhỏ hơn \$78, trong khi 50% số mặt hàng bán ra ở chi nhánh 2 có đơn giá nhỏ hơn \$85. Nói chung, ta nhận thấy có một sự dịch chuyển trong phân phối của chi nhánh 1 so với chi nhánh 2, khi đơn giá của các mặt hàng bán ra ở chi nhánh 1 có xu hướng thấp hơn so với chi nhánh 2.



Hình 2.5: Biểu đồ phân vị kép cho dữ liệu đơn giá từ hai chi nhánh của cửa hàng trực tuyến.

2.2.4.3 Biểu đồ tần suất

Biểu đồ tần suất (**histogram**) có lịch sử ít nhất một thế kỷ và được sử dụng rộng rãi. Từ "histos" có nghĩa là cột hoặc trụ, và "gram" có nghĩa là biểu đồ, vì vậy "histogram" là một biểu đồ gồm các cột đứng. Việc vẽ histogram là một phương pháp trực quan để tóm tắt phân phối của một thuộc tính cho trước, X. Tùy theo số lượng cột mong muốn trong biểu đồ, miền giá trị của X được chia thành một

tập hợp các khoảng con rời rạc liên tiếp. Thông thường, các khoảng có độ rộng bằng nhau. Ví dụ, thuộc tính *giá* với miền giá trị từ $1 đến $200 (làm tròn thành số nguyên) có thể được chia thành các khoảng con như $1-20$, $21-40$, $41-60$, v.v. Đối với mỗi khoảng con, một thanh dọc được vẽ với chiều cao tương ứng với tổng số lượng quan sát (hoặc tỷ lệ quan sát được) nằm trong khoảng đó.

Lưu ý rằng biểu đồ tần suất khác với một dạng biểu đồ phổ biến khác là **biểu đồ cột**. Biểu đồ cột sử dụng một tập hợp các thanh (thường có khoảng cách giữa các thanh) với trục X biểu diễn một tập hợp dữ liệu phân loại, chẳng hạn như *automobile_model* (mẫu xe hơi) hoặc *item_type* (loại mặt hàng). Chiều cao của thanh biểu thị số lượng của nhóm được xác định bởi danh mục đó. Ngược lại, biểu đồ tần suất thể hiện dữ liệu số, trong đó trục X đại diện cho một dãy giá trị liên tục được chia thành các khoảng. Biểu đồ tần suất được sử dụng để hiển thị phân phối dữ liệu (dọc theo trục X), trong khi biểu đồ cột được sử dụng để so sánh các danh mục. Biểu đồ tần suất có độ lệch, thể hiện xu hướng các quan sát tập trung nhiều ở đầu thấp hoặc đầu cao của trục X. Ngược lại, trục X của biểu đồ cột không có điểm đầu thấp hay cao, vì các nhãn trên trục này mang tính phân loại thay vì dạng số. Do đó, các cột trong biểu đồ cột có thể được sắp xếp lại nhưng các thanh trong biểu đồ tần suất thì không thể.

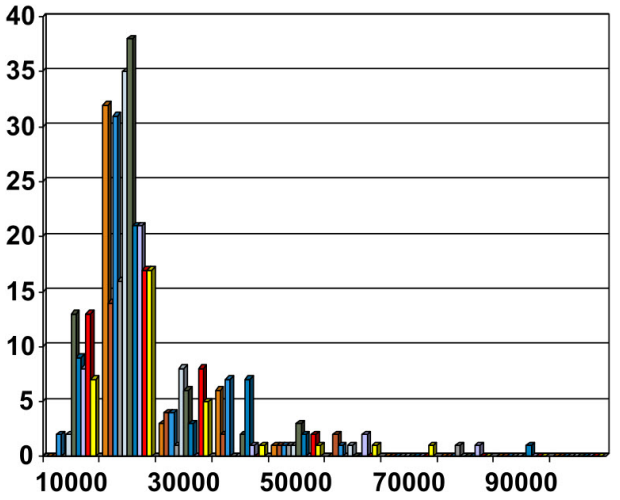
Ví dụ 22 (Biểu đồ tần suất)**.** Hình 2.6 hiển thị một biểu đồ tần suất về phân phối giải thưởng nghiên cứu cho một khu vực, trong đó các khoảng có độ rộng bằng nhau, mỗi khoảng đại diện cho mức tăng $1000. Tần số chính là số lượng giải thưởng nghiên cứu rơi vào từng khoảng giá trị này.

Mặc dù biểu đồ tần suất được sử dụng rộng rãi, nhưng chúng có thể không hiệu quả bằng các phương pháp như biểu đồ phân vị, biểu đồ phân vị kép hoặc biểu đồ hộp khi so sánh các nhóm quan sát đơn biến.

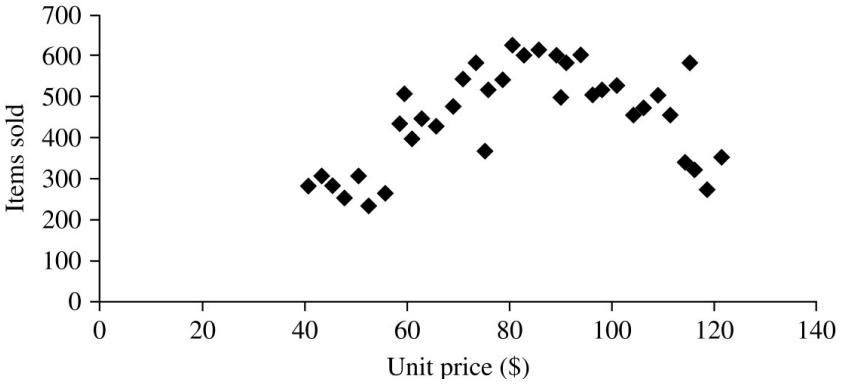
2.2.4.4 Biểu đồ phân tán và tương quan dữ liệu

Biểu đồ phân tán là một trong những phương pháp trực quan hiệu quả nhất để xác định xem liệu có mối quan hệ, mẫu, hoặc xu hướng nào giữa hai thuộc tính số hay không. Để xây dựng biểu đồ phân tán, mỗi cặp giá trị được xem như một cặp tọa độ trong không gian đại số (mặt phẳng tọa độ) và được vẽ dưới dạng các điểm trên mặt phẳng. Hình 2.7 cho thấy một biểu đồ phân tán cho tập dữ liệu trong Bảng 2.3.

Biểu đồ phân tán là công cụ hữu ích để có cái nhìn ban đầu về dữ liệu hai biến,



Hình 2.6: Một biểu đồ tần suất về phân phối giải thưởng nghiên cứu của một khu vực.



Hình 2.7: Một biểu đồ phân tán cho tập dữ liệu trong Bảng 2.3.

giúp nhận biết các cụm điểm và các điểm ngoại lệ, hoặc khám phá khả năng có mối tương quan giữa các thuộc tính. Hai thuộc tính X và Y được cho là **có tương quan** nếu việc biết giá trị của một thuộc tính cho phép dự đoán giá trị của thuộc tính kia với một độ chính xác nhất định. Các tương quan có thể là tương quan dương, tương quan âm hoặc không tương quan. Hình 2.8 cho thấy các ví dụ về tương quan dương và âm giữa hai thuộc tính.

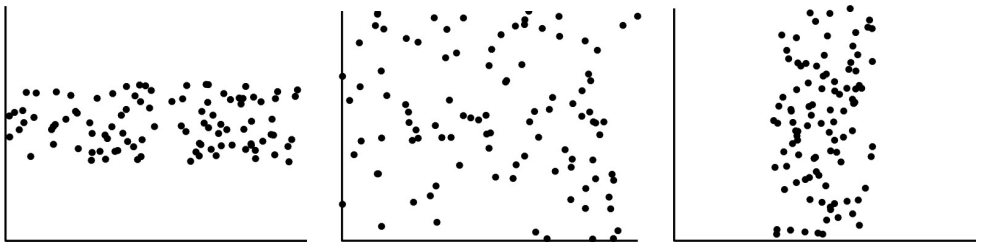
Nếu mẫu các điểm được vẽ có xu hướng nghiêng từ góc dưới bên trái lên góc trên bên phải, điều này có nghĩa là giá trị của X tăng khi giá trị của Y cũng tăng, gợi ý một *tương quan dương* (Hình 2.8a). Nếu mẫu các điểm có xu hướng nghiêng từ góc trên bên trái sang xuống góc dưới bên phải, nghĩa là giá trị của X tăng khi giá trị của Y giảm, gợi ý *một tương quan âm* (Hình 2.8b). Một đường hồi quy có thể

được vẽ để nghiên cứu mức độ tương quan giữa các biến. Các kiểm định thống kê về tương quan được giới thiệu trong Phụ lục A.

(a) (b)

Hình 2.8: Biểu đồ phân tán được sử dụng để tìm (a) tương quan dương hoặc (b) tương quan âm giữa các thuộc tính.

Hình 2.9 cho thấy ba trường hợp mà không có mối tương quan nào được quan sát giữa hai thuộc tính được vẽ cho mỗi tập dữ liệu.



Hình 2.9: Ba trường hợp không quan sát được mối tương quan giữa hai thuộc tính được vẽ trong mỗi tập dữ liệu.

Tóm lại, các mô tả thống kê cơ bản (ví dụ: các thước đo xu hướng trung tâm và độ phân tán) và các biểu đồ thống kê trực quan (ví dụ: biểu đồ phân vị, biểu đồ tần suất và biểu đồ phân tán) cung cấp những hiểu biết quý giá về hành vi tổng quát của dữ liệu. Chúng đặc biệt hữu ích cho việc làm sạch dữ liệu bằng cách giúp nhận diện nhiễu và các điểm ngoại lệ.

2.3 Độ tương đồng và khoảng cách

Trong các ứng dụng khai phá dữ liệu, chẳng hạn như phân cụm, phân tích ngoại lệ và phân loại theo phương pháp $k-$láng giềng gần nhất, chúng ta cần có

các cách để đánh giá mức độ giống nhau hoặc khác nhau giữa các đối tượng. Ví dụ, một cửa hàng có thể muốn tìm kiếm các cụm đối tượng *khách hàng*, tạo thành các nhóm khách hàng có đặc điểm tương tự (ví dụ: thu nhập, khu vực cư trú và tuổi tác tương đồng). Những thông tin như vậy có thể được sử dụng cho mục đích tiếp thị. Một **cụm** là tập hợp các đối tượng dữ liệu sao cho các đối tượng trong cùng một cụm có độ *tương đồng* cao và *khác biệt* với các đối tượng ở các cụm khác. Phân tích điểm ngoại lệ cũng sử dụng các kỹ thuật dựa trên phân cụm để nhận diện các đối tượng ngoại lai, tức là những đối tượng có độ khác biệt rất cao so với các đối tượng khác. Hiểu biết về mức độ tương đồng giữa các đối tượng cũng có thể được sử dụng trong các phương pháp phân loại theo phương pháp $k-$láng giềng gần nhất, trong đó một đối tượng cho trước (ví dụ: một *bệnh nhân*) được gán nhãn lớp (ví dụ, *chẩn đoán*) dựa trên mức độ tương đồng của nó với các đối tượng khác trong mô hình.

Phần này trình bày các thước đo độ tương đồng và độ khác biệt, được gọi chung là thước đo *sự gần gũi*. Một thước đo tương đồng của hai đối tượng i và j thường trả về giá trị 0 nếu các đối tượng hoàn toàn khác nhau. Giá trị tương đồng cao hơn biểu thị mức độ tương đồng lớn hơn giữa các đối tượng. (Thông thường, giá trị 1 biểu thị sự giống nhau hoàn toàn, nghĩa là các đối tượng là giống hệt nhau.) Ngược lại, một thước đo độ khác biệt trả về giá trị 0 nếu các đối tượng giống nhau (và do đó, hoàn toàn không có gì khác biệt). Giá trị độ khác biệt cao hơn cho thấy hai đối tượng có mức độ khác nhau lớn hơn.

Ở Mục 2.3.1, chúng ta trình bày hai cấu trúc dữ liệu thường được sử dụng trong các ứng dụng nói trên: *ma trận dữ liệu* (dùng để lưu trữ các đối tượng dữ liệu) và *ma trận độ khác biệt* (dùng để lưu trữ các giá trị độ khác biệt cho các cặp đối tượng). Ở đây, chúng ta thay đổi ký hiệu cho các đối tượng dữ liệu so với phần trước của chương, vì bây giờ chúng ta đang làm việc với các đối tượng được mô tả bởi nhiều thuộc tính. Từ đó, chúng ta sẽ thảo luận cách tính toán mức độ khác biệt của các đối tượng được mô tả bởi các thuộc tính danh định (Mục 2.3.2), các thuộc tính nhị phân (Mục 2.3.3), các thuộc tính số (Mục 2.3.4), các thuộc tính thứ tự (Mục 2.3.5), hoặc kết hợp của các loại thuộc tính này (Mục 2.3.6). Mục 2.3.7 cung cấp các thước đo tương đồng cho các véctơ dữ liệu dài và thưa, chẳng hạn như véctơ tần số từ đại diện cho các tài liệu trong truy xuất thông tin. Cuối cùng, Mục 2.3.8 thảo luận cách đo lường sự khác biệt giữa hai phân phối xác suất trên cùng một biến x, và giới thiệu một thước đo gọi là *độ phân kỳ Kullback–Leibler* (độ phân kỳ KL), được sử dụng phổ biến trong tài liệu khai phá dữ liệu.

Hiểu cách tính không tương đồng sẽ hữu ích trong việc nghiên cứu các thuộc

tính và cũng sẽ được tham khảo trong các chủ đề sau này về phân cụm (Chương 8 và 9), phân tích ngoại lệ (Chương 11) và phân loại theo phương pháp k – láng giềng gần nhất (Chương 6).

2.3.1 Ma trận dữ liệu và ma trận độ khác biệt

Trong Mục 2.2, chúng ta đã xem xét các cách nghiên cứu xu hướng trung tâm, độ phân tán và sự trải rộng của các giá trị quan sát được của một thuộc tính X. Ở đó, các đối tượng là một chiều, tức là được mô tả bởi một thuộc tính duy nhất. Trong mục này, chúng ta nói về các đối tượng được mô tả bởi *nhiều* thuộc tính, do đó cần thay đổi ký hiệu. Giả sử rằng chúng ta có n đối tượng (ví dụ: người, mặt hàng, hoặc khóa học) được mô tả bởi p thuộc tính (còn gọi là *phép đo* hoặc *đặc trưng*, chẳng hạn như tuổi, chiều cao, cân nặng hoặc giới tính). Các đối tượng được biểu diễn bởi $x^{(1)} = \left(x_{11}, x_{12}, \dots, x_{1p}\right)^\top$, $x^{(2)} = \left(x_{21}, x_{22}, \dots, x_{2p}\right)^\top$, ..., trong đó x_{ij} là giá trị của đối tượng $x^{(i)}$ đối với thuộc tính thứ j. Để ngắn gọn, sau đây chúng ta sẽ gọi đối tượng $x^{(i)}$ là đối tượng i. Các đối tượng này có thể là các bộ dữ liệu trong một cơ sở dữ liệu quan hệ và còn được gọi là các *mẫu dữ liệu* hoặc *véctơ đặc trưng*.

Các thuật toán chính về phân cụm và k – láng giềng gần nhất dựa trên bộ nhớ thường hoạt động trên một trong hai cấu trúc dữ liệu sau:

- **Ma trận dữ liệu** (hay *cấu trúc đối tượng – thuộc tính*): Cấu trúc này lưu trữ n đối tượng dữ liệu dưới dạng một bảng quan hệ hoặc ma trận có kích thước $n \times p$ (n đối tượng $\times p$ thuộc tính):

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ & \dots\dots\dots & & & \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ & \dots\dots\dots & & & \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix} \tag{2.13}$$

  Mỗi hàng tương ứng với một đối tượng. Trong ký hiệu của chúng ta, có thể dùng f để đánh chỉ số qua các thuộc tính từ 1 đến p.

- Ma trận độ khác biệt (hay *cấu trúc đối tượng – đối tượng*): Cấu trúc này lưu trữ một tập hợp các giá trị gần gũi có sẵn cho tất cả các cặp của n đối tượng.

Nó thường được biểu diễn dưới dạng một bảng kích thước $n \times n$:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ & \cdots & \cdots & \cdots & \\ d(2,1) & d(n,2) & d(n,3) & \ldots & 0 \end{bmatrix} \tag{2.14}$$

trong đó $d(i,j)$ là **độ khác biệt** hoặc "độ không tương đồng" được đo giữa đối tượng i và j. Nói chung, $d(i,j)$ là một số không âm và sẽ gần bằng 0 khi đối tượng i và j có mức độ giống nhau rất cao hoặc "gần" nhau, và tăng lên khi chúng ngày càng khác nhau. Lưu ý rằng $d(i,i) = 0$; tức là, sự khác biệt giữa một đối tượng và chính nó là 0. Hơn nữa, $d(i,j) = d(j,i)$. (Vì bảng đối xứng, nên chúng ta không hiển thị các giá trị $d(j,i)$.)

Các thước đo độ khác biệt thường được dùng để biểu diễn mức độ tương đồng giữa các đối tượng theo một hàm số của thước đo độ khác biệt. Ví dụ, với dữ liệu danh định, ta có thể định nghĩa:

$$\text{sim}(i,j) = 1 - d(i,j), \tag{2.15}$$

trong đó $\text{sim}(i,j)$ là mức độ tương đồng giữa các đối tượng i và j.

Một ma trận dữ liệu được tạo thành từ hai thực thể (hay, hai thành phần, "hai thứ"), đó là các hàng (cho các đối tượng) và các cột (cho các thuộc tính). Do đó, ma trận dữ liệu thường được gọi là ma trận **hai mô hình** (two – mode matrix). Trong khi đó, ma trận độ khác biệt chứa một loại thực thể duy nhất (các giá trị độ khác biệt) và do đó được gọi là ma trận **một mô hình** (one – mode). Nhiều thuật toán phân cụm và k – láng giềng gần nhất hoạt động trên ma trận độ khác biệt. Dữ liệu ở dạng ma trận dữ liệu có thể được chuyển đổi thành ma trận độ khác biệt trước khi áp dụng các thuật toán này.

2.3.2 Thước đo gần gũi cho thuộc tính danh định

Một thuộc tính danh định có thể nhận được hai hoặc nhiều trạng thái (xem Mục 2.1.1). Ví dụ, thuộc tính "*map_color*" (màu bản đồ) là thuộc tính danh định có thể có, giả sử, năm trạng thái: *đỏ*, *vàng*, *xanh lá*, *hồng*, và *xanh dương*.

Giả sử số trạng thái của một thuộc tính danh định là M. Các trạng thái có thể được ký hiệu bằng chữ cái, ký hiệu hoặc một tập hợp các số nguyên, chẳng hạn

như 1, 2, ... , M. Lưu ý rằng các số nguyên này chỉ được sử dụng để xử lý dữ liệu và không biểu thị bất kỳ thứ tự cụ thể nào.

"*Làm thế nào để tính độ khác biệt giữa các đối tượng được mô tả bởi các thuộc tính danh định*?" Độ khác biệt giữa hai đối tượng i và j có thể được tính dựa trên tỷ lệ số lượng không khớp như sau:

$$d(i,j) = \frac{p-m}{p}, \tag{2.16}$$

trong đó m là số lượng các thuộc tính mà đối tượng i và j có cùng trạng thái, và p là tổng số thuộc tính mô tả các đối tượng. Các trọng số có thể được gán để tăng tác động của m hoặc để gán trọng số lớn hơn cho những thuộc tính có số trạng thái nhiều hơn.

Ví dụ 23 (Độ khác biệt giữa các thuộc tính danh định)**.** Giả sử chúng ta có dữ liệu mẫu trong Bảng 2.4, chỉ có thông tin *object−identifier* (nhận dạng đối tượng) và thuộc tính (test−1) (là thuộc tính danh định). (Chúng ta sẽ sử dụng *test−2* và *test−3* trong các ví dụ sau.) Hãy tính ma trận độ khác biệt theo công thức (2.14), tức là ma trận

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Vì ở đây chỉ có một thuộc tính danh định, *test-1*, ta đặt $p = 1$ trong công thức (2.16) nên $d(i,j)$ sẽ bằng 0 nếu các đối tượng i và j có cùng giá trị, và bằng 1 nếu chúng khác nhau. Như vậy, ta được

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Từ đó, ta thấy rằng tất cả các đối tượng đều khác biệt, ngoại trừ đối tượng 1 và đối tượng 4 (tức là, $d(4,1) = 0$).

Ngoài ra, ta có thể tính độ tương đồng theo công thức:

$$\text{sim}(i,j) = 1 - d(i,j) = \frac{m}{p}. \tag{2.17}$$

Object Identifier	**Test-1** (danh định)	**Test-2** (thứ tự)	**Test-3** (số)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Bảng 2.4: Một bảng dữ liệu mẫu chứa hỗn hợp các thuộc tính.

Độ gần gũi giữa các đối tượng được mô tả bởi các thuộc tính danh định cũng có thể được tính bằng cách sử dụng một phương pháp mã hóa thay thế. Các thuộc tính danh định có thể được mã hóa thành các thuộc tính nhị phân bất đối xứng bằng cách tạo một thuộc tính nhị phân mới cho mỗi trạng thái trong M trạng thái. Đối với một đối tượng có một giá trị trạng thái nhất định, thuộc tính nhị phân đại diện cho trạng thái đó được gán giá trị 1, trong khi các thuộc tính nhị phân còn lại được gán giá trị 0. Ví dụ, để mã hóa thuộc tính danh định *map_color*, có thể tạo ra một thuộc tính nhị phân cho mỗi trong số năm màu đã liệt kê. Với một đối tượng có màu *vàng*, thuộc tính *yellow* được gán là 1, trong khi bốn thuộc tính còn lại được gán là 0. Các thước đo gần gũi cho dạng mã hóa này có thể được tính bằng các phương pháp được bàn ở phần tiếp theo.

2.3.3 Thước đo gần gũi cho thuộc tính nhị phân

Hãy cùng xem cách tính toán độ khác biệt và độ tương đồng cho các đối tượng được mô tả bởi các *thuộc tính nhị phân đối xứng* hoặc *bất đối xứng*.

Nhớ rằng một thuộc tính nhị phân chỉ có hai trạng thái: 0 và 1, trong đó 0 biểu thị rằng thuộc tính không có mặt, và 1 biểu thị rằng thuộc tính có mặt (xem Mục 2.1.2). Ví dụ, với thuộc tính *smoker* (hút thuốc) mô tả một bệnh nhân, giá trị 1 cho biết bệnh nhân có hút thuốc, còn 0 cho biết bệnh nhân không hút thuốc. Việc xử lý các thuộc tính nhị phân như các thuộc tính số khác có thể dẫn đến những kết quả sai lệch, do đó cần các phương pháp riêng cho dữ liệu nhị phân để tính toán độ khác biệt.

"*Vậy làm thế nào để tính độ khác biệt giữa hai đối tượng dựa trên các thuộc tính nhị phân?*" Một cách tiếp cận là tính toán ma trận không tương đồng từ dữ liệu nhị phân cho trước. Nếu ta coi tất cả các thuộc tính nhị phân có cùng trọng số, ta có bảng tiếp liên 2×2 như Bảng 2.5, trong đó q là số thuộc tính mà đối tượng i và j đều có giá trị 1, r là số thuộc tính mà đối tượng i có giá trị 1 nhưng đối tượng j có

giá trị 0, s là số thuộc tính mà đối tượng i có giá trị 0 nhưng đối tượng j có giá trị 1, và t là số thuộc tính mà cả hai đối tượng i và j đều có giá trị 0. Tổng số thuộc tính là p, với $p = q + r + s + t$.

		Đối tượng *j*		
		1	0	Tổng
Đối tượng *i*	1	q	r	$q + r$
	0	s	t	$s + t$
	Tổng	$q + s$	$r + t$	p

Bảng 2.5: Bảng tiếp liên cho các thuộc tính nhị phân.

Nhắc lại rằng với các thuộc tính nhị phân đối xứng, mỗi trạng thái đều có giá trị như nhau. Độ khác biệt dựa trên các thuộc tính nhị phân đối xứng được gọi là **độ khác biệt nhị phân đối xứng**. Nếu các đối tượng i và j được mô tả bởi các thuộc tính nhị phân đối xứng, thì độ khác biệt giữa i và j được tính theo công thức:

$$d(i, j) = \frac{r + s}{q + r + s + t}. \tag{2.18}$$

Đối với thuộc tính nhị phân bất đối xứng, các trạng thái không có tầm quan trọng như nhau (ví dụ: kết quả *dương tính* (1) của xét nghiệm bệnh là quan trọng hơn kết quả âm tính (0)), các thuộc tính nhị phân thường được coi là "một mặt" (monary, có một trạng thái). Trong trường hợp này, sự phù hợp của hai giá trị 1 (phù hợp dương) được xem là quan trọng hơn so với sự phù hợp của hai giá trị 0 (phù hợp âm). Vì vậy, số lượng các giá trị 0 khớp, t, thường bị bỏ qua trong tính toán. Độ khác biệt cho thuộc tính nhị phân bất đối xứng, gọi là **độ khác biệt nhị phân bất đối xứng**, được tính bằng:

$$d(i, j) = \frac{r + s}{q + r + s}. \tag{2.19}$$

Ngược lại, ta có thể tính mức độ tương đồng giữa hai đối tượng dựa trên ý niệm về tương đồng thay vì sự khác biệt. Ví dụ, **độ tương đồng nhị phân bất đối xứng** giữa các đối tượng i và j được tính như sau:

$$\text{sim}(i, j) = \frac{q}{q + r + s} = 1 - d(i, j). \tag{2.20}$$

Hệ số $\text{sim}(i, j)$ của công thức (2.20) được gọi là **hệ số Jaccard**, là thước đo được tham khảo phổ biến trong tài liệu khai phá dữ liệu.

Khi trong cùng một tập dữ liệu xuất hiện cả thuộc tính nhị phân đối xứng và bất đối xứng, ta có thể áp dụng phương pháp cho thuộc tính hỗn hợp như được mô tả trong Mục 2.3.6.

Ví dụ 24 (Độ không tương đồng giữa các thuộc tính nhị phân)**.** Giả sử bảng hồ sơ bệnh nhân (Bảng 2.6) chứa các thuộc tính: *Name* (tên), *Gender* (giới tính), *Fever* (sốt), *Cough* (ho), *Test*-1, *Test*-2, *Test*-3, và *Test*-4. Trong đó, thuộc tính *Name* là định danh đối tượng, *Gender* là thuộc tính nhị phân đối xứng, và các thuộc tính còn lại là thuộc tính nhị phân bất đối xứng.

Name	**Gender**	**Fever**	**Cough**	**Test**-1	**Test**-2	**Test**-3	**Test**-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Bảng 2.6: Bảng quan hệ mô tả các bệnh nhân bằng các thuộc tính nhị phân.

Đối với các thuộc tính nhị phân bất đối xứng, ta quy ước gán giá trị 1 cho các giá trị *Y* (*yes*) và *P* (*positive*), còn giá trị *N* (*no* hoặc *negative*) được gán là 0. Giả sử khoảng cách giữa các đối tượng (bệnh nhân) được tính dựa chỉ trên các thuộc tính nhị phân bất đối xứng. Theo công thức (2.19) cho thuộc tính nhị phân bất đối xứng, khoảng cách giữa các cặp bệnh nhân trong ba bệnh nhân Jack, Mary, và Jim là

$$d(\text{Jack}, \text{Jim}) = \frac{1+1}{1+1+1} = \frac{2}{3} \approx 0.67.$$
$$d(\text{Jack}, \text{Mary}) = \frac{0+1}{2+0+1} = \frac{1}{3} \approx 0.33.$$
$$d(\text{Jim}, \text{Mary}) = \frac{1+2}{1+1+2} = \frac{3}{4} = 0.75.$$

Những giá trị tính được cho thấy Jim và Mary có độ khác biệt cao nhất (0.75), gợi ý rằng khả năng họ mắc cùng một bệnh là thấp. Trong số ba bệnh nhân, Jack và Mary có độ khác biệt thấp nhất (0.33), cho thấy họ có khả năng mắc bệnh tương tự cao hơn.

2.3.4 Độ khác biệt của dữ liệu số: khoảng cách Minkowski

Trong phần này, chúng ta sẽ mô tả các thước đo khoảng cách thường được sử dụng để tính toán độ khác biệt của các đối tượng được mô tả bởi các thuộc tính số. Các thước đo này bao gồm *khoảng cách Euclid*, *khoảng cách Manhattan* và *khoảng cách Minkowski*.

Trong một số trường hợp, dữ liệu được chuẩn hóa trước khi áp dụng các phép tính khoảng cách. Điều này liên quan đến việc biến đổi dữ liệu để nằm trong một khoảng giá trị nhỏ hơn hoặc phổ biến, chẳng hạn như $[-1.0, 1.0]$ hoặc $[0.0, 1.0]$. Ví dụ, thuộc tính *chiều cao* có thể được đo bằng mét hoặc inch. Nói chung, biểu diễn một thuộc tính bằng các đơn vị nhỏ hơn sẽ dẫn đến một miền giá trị lớn hơn của thuộc tính đó, từ đó có xu hướng làm cho thuộc tính này có "trọng số" ảnh hưởng lớn hơn. Việc chuẩn hóa dữ liệu cố gắng gán cho tất cả các thuộc tính trọng số bằng nhau. Tuy nhiên, điều này có thể có ích hoặc không tùy vào ứng dụng cụ thể. Các phương pháp chuẩn hóa dữ liệu được bàn luận chi tiết ở Mục 2.5 về biến đổi dữ liệu.

Khoảng cách Euclid (tức khoảng cách đường thẳng hay "theo đường chim bay") là thước đo phổ biến nhất. Giả sử đối tượng $i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^\top$ và đối tượng $j = \left(x_{j1}, x_{j2}, \ldots, x_{jp}\right)^\top$ được mô tả bởi p thuộc tính số. Khoảng cách Euclid giữa hai đối tượng i và j được định nghĩa là:

$$d(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2}. \tag{2.21}$$

Khoảng cách Manhattan (hay khoảng cách "theo khối" hoặc "theo dãy phố") được đặt tên như vậy bởi vì nó thể hiện khoảng cách tính theo số khối giữa hai điểm trong thành phố (ví dụ: 3 khối về phía dưới và 2 khối sang bên, tổng cộng là 5 khối). Nó được định nghĩa là:

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{ip} - x_{jp}\right|. \tag{2.22}$$

Cả khoảng cách Euclid và Manhattan đều thỏa mãn các tính chất toán học sau:

Không âm: $d(i,j) \geq 0$. Khoảng cách luôn là một số không âm.

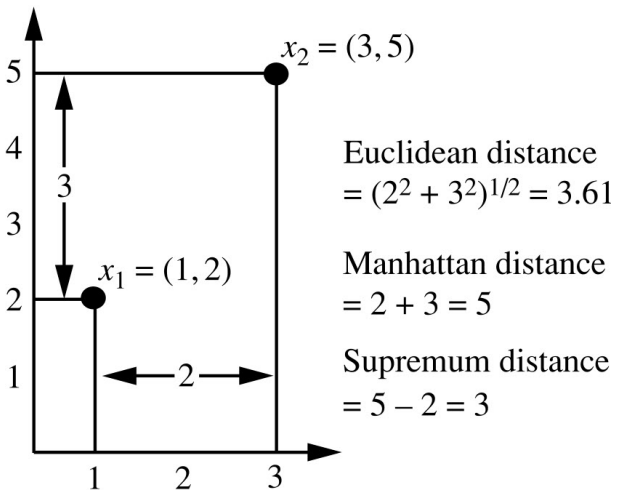
Định danh của các đối tượng không thể phân biệt: $d(i,i) = 0$. Khoảng cách của một đối tượng so với chính nó là 0.

Đối xứng $d(i,j) = d(j,i)$. Khoảng cách là hàm đối xứng.

Bất đẳng thức tam giác $d(i,j) \leq d(i,k) + d(k,j)$. Việc đi thẳng từ đối tượng i đến đối tượng j không vượt quá việc đi qua bất kỳ đối tượng k nào.

Một thước đo thoả mãn những điều kiện này được gọi là **metric**. Lưu ý rằng tính chất không âm thường được suy ra từ ba tính chất còn lại.

Ví dụ 25 (Khoảng cách Euclid và Manhattan)**.** Giả sử $x^{(1)} = (1,2)$ và $x^{(2)} = (3,5)$ là hai đối tượng như minh họa trong Hình 2.10. Khoảng cách Euclid giữa hai đối tượng là $\sqrt{(3-1)^2 + (5-2)^2} = \sqrt{2^2 + 3^2} = \sqrt{13} \approx 3.61$. Khoảng cách Manhattan giữa chúng là: $|3-1| + |5-2| = 2 + 3 = 5$.



Hình 2.10: Khoảng cách Euclid, khoảng cách Manhattan, và khoảng cách supremum.

Khoảng cách Minkowski là một mở rộng của khoảng cách Euclid và Manhattan. Nó được định nghĩa là:

$$d(i,j) = \sqrt[h]{\left(x_{i1} - x_{j1}\right)^h + \left(x_{i2} - x_{j2}\right)^h + \cdots + \left(x_{ip} - x_{jp}\right)^h}, \tag{2.23}$$

trong đó h là một số thực sao cho $h \geq 1$. (Khoảng cách này còn được gọi là **chuẩn** L_p trong một số tài liệu, trong đó ký hiệu p tương đương với h theo ký hiệu của chúng ta. Chúng ta giữ p làm số thuộc tính để thống nhất với phần còn lại của chương.) Khi $h = 1$ (tức là chuẩn L_1 norm), khoảng cách Minkowski trở thành khoảng cách Manhattan; khi $h = 2$ (tức là chuẩn L_2), nó trở thành khoảng cách Euclid.

Khoảng cách supremum (còn được gọi là chuẩn $L_{\max}$, chuẩn L_∞, hay **khoảng cách Chebyshev**) là một mở rộng của khoảng cách Minkowski khi $h \to \infty$. Để tính toán khoảng cách này, ta tìm thuộc tính f cho thấy sự khác biệt lớn nhất về giá trị giữa hai đối tượng. Sự khác biệt này được gọi là khoảng cách supremum, được định nghĩa chính thức như sau:

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{1\le f\le p} |x_{if} - x_{jf}| . \tag{2.24}$$

Chuẩn L_∞ còn được gọi là *chuẩn đơn trị*.

Ví dụ 26 (Khoảng cách supremum)**.** Sử dụng hai đối tượng $x^{(1)} = (1,2)$ và $x^{(2)} = (3,5)$ như trong Hình 2.10. Thuộc tính thứ hai cho thấy sự khác biệt lớn nhất giữa giá trị của hai đối tượng. Cụ thể, $\max\{|3-1|, |5-2|\} = 3$. Đây chính là khoảng cách supremum giữa hai đối tượng.

Nếu mỗi thuộc tính được gán một trọng số theo mức độ quan trọng được đánh giá, thì **khoảng cách Euclid có trọng số** có thể được tính bằng công thức:

$$d(i,j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \cdots + w_p|x_{ip} - x_{jp}|^2}. \tag{2.25}$$

Trọng số cũng có thể được áp dụng cho các thước đo khoảng cách khác.

2.3.5 Thước đo gần gũi cho thuộc tính thứ tự

Các giá trị của một thuộc tính thứ tự có thứ tự hoặc xếp hạng có ý nghĩa, nhưng khoảng cách giữa các giá trị liên tiếp không được biết rõ (xem Mục 2.1.3). Ví dụ, đối với thuộc tính *kích cỡ*, chuỗi *small*, *medium*, *large* thể hiện thứ tự. Các thuộc tính thứ tự cũng có thể được thu được từ việc rời rạc hóa các thuộc tính số bằng cách chia khoảng giá trị thành một số hữu hạn các danh mục. Những danh mục này được sắp xếp theo thứ tự xếp hạng. Tức là, khoảng giá trị của một thuộc tính số có thể được ánh xạ sang một thuộc tính thứ tự f với M_f trạng thái. Ví dụ, khoảng giá trị của thuộc tính theo thang đo khoảng *nhiệt độ* (theo độ C) có thể được chia thành các trạng thái: -30 đến -10, -10 đến 10, và 10 đến 30, tương ứng với các danh mục *nhiệt độ lạnh*, *nhiệt độ vừa phải*, và *nhiệt độ ấm*. Gọi M_f là số trạng thái khả dĩ của một thuộc tính thứ tự. Những trạng thái có thứ tự này định nghĩa xếp hạng $1, \ldots, M_f$.

"*Làm thế nào để xử lý các thuộc tính thứ tự?*" Việc xử lý các thuộc tính thứ tự khá tương tự với việc xử lý các thuộc tính số khi tính toán độ khác biệt giữa các đối tượng. Giả sử f là một thuộc tính trong tập các thuộc tính thứ tự mô tả n đối tượng. Quá trình tính độ khác biệt đối với f gồm các bước sau:

1) Chuyển đổi giá trị thành xếp hạng: Giá trị của thuộc tính f đối với đối tượng thứ i là x_{if}, và f có M_f trạng thái có thứ tự, biểu diễn xếp hạng từ $1, \dots, M_f$. Thay thế mỗi x_{if} bằng xếp hạng tương ứng của nó, $r_{if} \in \{1, \dots, M_f\}$.

2) Chuẩn hóa xếp hạng: Vì mỗi thuộc tính thứ tự có thể có số trạng thái khác nhau, nên thường cần phải ánh xạ khoảng giá trị của mỗi thuộc tính về khoảng $[0.0, 1.0]$ để tất cả các thuộc tính có trọng số bằng nhau. Ta thực hiện việc chuẩn hóa bằng cách thay thế xếp hạng r_{if} của đối tượng thứ i trong thuộc tính thứ f bởi:
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}. \tag{2.26}$$

3) Tính độ khác biệt: Sau đó, ta có thể tính độ khác biệt sử dụng bất kỳ thước đo khoảng cách nào đã được mô tả ở Mục 2.3.4 cho các thuộc tính số, sử dụng z_{if} làm biểu diễn giá trị của thuộc tính f đối với đối tượng thứ i.

Ví dụ 27 (Độ khác biệt giữa các thuộc tính thứ tự)**.** Giả sử ta có dữ liệu mẫu được trình bày trong Bảng 2.4, nhưng lần này chỉ có thông tin định danh đối tượng *object−identifier* và thuộc tính thứ tự liên tục, *test−2*. Có ba trạng thái cho *test−2*: *fair*, *good*, và *excellent*, tức $M_f = 3$. Tại bước 1, nếu thay thế mỗi giá trị của *test−2* bằng xếp hạng của nó, bốn đối tượng sẽ được gán các xếp hạng lần lượt là: 3, 1, 2, và 3. Bước 2 chuẩn hóa các xếp hạng bằng cách ánh xạ xếp hạng 1 thành 0.0, xếp hạng 2 thành 0.5, và xếp hạng 3 thành 1.0. Với bước 3, sử dụng ví dụ khoảng cách Euclid (được định nghĩa ở công thức (2.21)) để tính toán độ khác biệt, ta được ma trận khác biệt như sau:

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Như vậy, đối tượng 1 và 2 có độ khác biệt cao nhất (ví dụ, $d(2,1) = 1.0$), cũng như đối tượng 2 và 4 ($d(4,2) = 1.0$). Điều này có ý nghĩa trực quan, vì đối tượng 1 và 4 đều có giá trị *excellent* (xếp hạng 3), trong khi đối tượng 2 có giá trị *fair* (xếp hạng 1), nằm ở đầu dưới của miền giá trị của *test−2*.

Các giá trị tương đồng cho thuộc tính thứ tự có thể được tính từ độ khác biệt bằng công thức $\text{sim}(i, j) = 1 - d(i, j)$.

2.3.6 Độ không tương đồng cho thuộc tính hỗn hợp

Các Mục 2.3.2 đến 2.3.5 đã thảo luận cách tính độ khác biệt giữa các đối tượng được mô tả bởi các thuộc tính cùng loại, trong đó các loại này có thể là *danh định*, *nhị phân đối xứng*, *nhị phân bất đối xứng*, *số*, hoặc *thứ tự*. Tuy nhiên, trong nhiều cơ sở dữ liệu thực tế, các đối tượng được mô tả bằng sự *kết hợp* của nhiều loại thuộc tính. Nói chung, một cơ sở dữ liệu có thể chứa tất cả các loại thuộc tính này.

"Vậy làm thế nào để tính độ khác biệt giữa các đối tượng có thuộc tính hỗn hợp?" Một cách tiếp cận là gom nhóm các thuộc tính theo loại, thực hiện phân tích khai phá (ví dụ: phân cụm) riêng cho mỗi loại. Phương pháp này khả thi nếu các phân tích đó cho ra kết quả tương thích. Tuy nhiên, trong các ứng dụng thực tế, khả năng một phân tích riêng biệt cho từng loại thuộc tính cho ra kết quả tương thích là khá thấp.

Một phương pháp được ưu tiên hơn là xử lý tất cả các loại thuộc tính cùng một lúc, thực hiện một phân tích duy nhất. Một kỹ thuật như vậy kết hợp các thuộc tính khác nhau vào một ma trận độ khác biệt đơn, đưa tất cả các thuộc tính có ý nghĩa về cùng một thang đo trên khoảng [0.0, 1.0].

Giả sử tập dữ liệu chứa p thuộc tính hỗn hợp. Độ không tương đồng $d(i, j)$ giữa các đối tượng i và j được định nghĩa là:

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}, \tag{2.27}$$

trong đó chỉ báo $\delta_{ij}^{(f)} = 0$ nếu hoặc là (1) x_{if} hoặc x_{jf} bị thiếu (tức là không có phép đo của thuộc tính f cho đối tượng i hoặc đối tượng j), hoặc (2) $x_{if} = x_{jf} = 0$ và thuộc tính f là nhị phân bất đối xứng; ngược lại, $\delta_{ij}^{(f)} = 1$. Đóng góp của thuộc tính f vào độ khác biệt giữa i và j (tức $d_{ij}^{(f)}$) được tính tùy theo loại của thuộc tính:

- Nếu f là thuộc tính số: $d_{ij}^{(f)} = \dfrac{|x_{if} - x_{jf}|}{\max_f - \min_f}$, trong đó $\max_f$ và $\min_f$ lần lượt là giá trị lớn nhất và nhỏ nhất của thuộc tính f.

- Nếu f là thuộc tính danh định hoặc nhị phân: $d_{ij}^{(f)} = 0$ nếu $x_{if} = x_{jf}$; ngược lại, $d_{ij}^{(f)} = 1$.

- Nếu f là thuộc tính thứ tự: ta tính xếp hạng r_{if} và sau đó chuẩn hóa thành $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, và coi z_{if} như là giá trị số của thuộc tính f.

Những bước này giống hệt với những gì đã trình bày cho từng loại thuộc tính riêng lẻ. Sự khác biệt duy nhất là đối với thuộc tính số, ta cần chuẩn hóa để các giá trị được ánh xạ về khoảng [0.0, 1.0]. Nhờ đó, ta có thể tính được độ khác biệt giữa các đối tượng ngay cả khi các thuộc tính mô tả các đối tượng có các loại khác nhau.

Ví dụ 28 (Độ khác biệt giữa các đối tượng có thuộc tính hỗn hợp)**.** Hãy tính ma trận độ khác biệt cho các đối tượng trong Bảng 2.4. Lần này, chúng ta sẽ xem xét *tất cả* các thuộc tính, vốn thuộc các loại khác nhau. Trong các Ví dụ 23 và 27, chúng ta đã tính toán ma trận độ khác biệt cho từng thuộc tính riêng lẻ. Quy trình xử lý cho *test−1* (danh định) và *test−2* (thứ tự) giống như đã nêu ở trên khi xử lý các thuộc tính hỗn hợp. Do đó, ta có thể sử dụng các ma trận độ khác biệt thu được cho *test−1* và *test−2* sau này khi tính công thức (2.27). Trước tiên, ta cần tính ma trận độ khác biệt cho thuộc tính thứ ba, *test−3* (là thuộc tính số). Cụ thể, ta phải tính $d_{ij}^{(3)}$. Theo cách xử lý đối với thuộc tính số, giả sử ta có $\max_{\text{test}-3} = 64$ và $\min_{\text{test}-3} = 22$. Hiệu giữa hai giá trị này được sử dụng trong công thức (2.27) để chuẩn hóa các giá trị trong ma trận độ khác biệt. Ma trận độ khác biệt thu được cho *test−3* là:

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Chúng ta có thể sử dụng các ma trận độ khác biệt của ba thuộc tính trong việc tính công thức (2.27). Ở đây, chỉ báo $\delta_{ij}^{(f)} = 1$ cho mỗi trong ba thuộc tính f. Ví dụ, ta có $d(3,1) = \frac{1(1) + 1(0.50) + 1(0.45)}{3} = 0.65$. Ma trận độ khác biệt thu được cho dữ liệu được mô tả bởi ba thuộc tính hỗn hợp là:

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Từ Bảng 2.4, ta có thể dễ dàng nhận xét rằng các đối tượng 1 và 4 có độ tương đồng cao nhất, dựa trên các giá trị của *test−1* và *test−2*. Điều này được xác nhận

bởi ma trận độ khác biệt, trong đó d (4, 1) là giá trị thấp nhất đối với bất kỳ cặp đối tượng nào khác. Tương tự, ma trận cho thấy rằng đối tượng 1 và đối tượng 2 có độ tương đồng thấp nhất.

2.3.7 Độ tương đồng cosine

Độ tương đồng cosine đo lường mức độ tương đồng giữa hai véctơ trong một không gian có tích vô hướng. Nó được tính bằng cosine của góc giữa hai véctơ và cho biết liệu hai véctơ có hướng chung gần như nhau hay không. Phương pháp này thường được sử dụng để đo mức độ tương đồng giữa các tài liệu trong phân tích văn bản.

Một tài liệu có thể được biểu diễn bằng hàng nghìn thuộc tính, mỗi thuộc tính ghi nhận tần số xuất hiện của một từ (hoặc cụm từ) nhất định trong tài liệu. Do đó, mỗi tài liệu là một đối tượng được biểu diễn dưới dạng *véctơ tần số từ*. Ví dụ, trong Bảng 2.7, ta thấy *Document1* có 5 lần xuất hiện từ *team*, trong khi *hockey* xuất hiện 3 lần. Từ *coach* không xuất hiện trong tài liệu (được biểu thị bởi giá trị 0). Những dữ liệu như vậy có thể rất bất đối xứng.

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Bảng 2.7: Véc tơ tài liệu hay véctơ tần số từ.

Các véctơ tần số từ thường rất dài và **thưa** (có rất nhiều giá trị 0). Các ứng dụng sử dụng cấu trúc này bao gồm truy xuất thông tin, phân cụm tài liệu văn bản, và phân tích dữ liệu sinh học. Các thước đo khoảng cách truyền thống đã được nghiên cứu trong chương này không hoạt động tốt với những dữ liệu số thưa như vậy. Ví dụ, hai véctơ tần số từ nói chung có thể có rất nhiều giá trị 0, có nghĩa là các tài liệu tương ứng không có nhiều từ chung, nhưng điều này không làm cho chúng trở nên tương đồng. Chúng ta cần một thước đo tập trung vào các từ mà hai tài liệu *cùng có*, và tần số xuất hiện của các từ đó. Nói cách khác, chúng ta cần một thước đo cho dữ liệu số mà bỏ qua các phép khớp với giá trị 0.

Độ tương đồng cosine là một thước đo tương đồng có thể được sử dụng để so sánh các tài liệu hoặc để xếp hạng các tài liệu theo một véctơ các từ truy vấn cho trước. Giả sử $\boldsymbol{x}$ và $\boldsymbol{y}$ là hai véctơ cần so sánh. Sử dụng thước đo cosine làm

hàm tương đồng, ta có:

$$\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \ \|\boldsymbol{y}\|}, \tag{2.28}$$

trong đó $\|\boldsymbol{x}\|$ là chuẩn Euclid của véctơ $\boldsymbol{x} = \left(x_1, x_2, \ldots, x_p\right)^\top$, được định nghĩa là $\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$. Khái niệm này tương đương với chiều dài của véctơ. Tương tự, $\|\boldsymbol{y}\|$ là chuẩn Euclid của véctơ $\boldsymbol{y}$. Thước đo này tính cosine của góc giữa véctơ $\boldsymbol{x}$ và $\boldsymbol{y}$. Một giá trị cosine bằng 0 nghĩa là hai véctơ vuông góc với nhau và không có điểm chung. Giá trị cosine càng gần 1 thì góc giữa hai véctơ càng nhỏ và mức độ giống nhau càng cao. Lưu ý rằng vì thước đo tương đồng cosine không thỏa mãn tất cả các tính chất của các thước đo metric (xem Mục 2.3.4), nên nó được gọi là *thước đo phi metric*.

Ví dụ 29 (Độ tương đồng cosine giữa hai véctơ tần số từ)**.** Giả sử x và y là hai véctơ tần số từ đầu tiên trong Bảng 2.7, cụ thể: $\boldsymbol{x} = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)^\top$, và $\boldsymbol{y} = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)^\top$. $\boldsymbol{x}$ và $\boldsymbol{y}$ tương đồng với nhau như thế nào? Sử dụng công thức (2.28) để tính độ tương đồng cosine giữa hai véctơ, ta có

$$\begin{aligned}
\boldsymbol{x} \cdot \boldsymbol{y} &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 \\
&\quad + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25. \\
\|\boldsymbol{x}\| &= \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} \approx 6.48. \\
\|\boldsymbol{y}\| &= \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} \approx 4.12. \\
\text{sim}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{25}{6.48 \times 4.12} \approx 0.94.
\end{aligned}$$

Do đó, nếu sử dụng thước đo tương đồng cosine để so sánh các tài liệu, chúng sẽ được coi là khá tương đồng.

Khi các thuộc tính có giá trị nhị phân, hàm tương đồng cosine có thể được diễn giải theo khái niệm về các đặc trưng hay thuộc tính chung. Giả sử một đối tượng $\boldsymbol{x}$ có thuộc tính thứ i nếu $x_i = 1$. Khi đó, $\boldsymbol{x} \cdot \boldsymbol{y}$ là số lượng các thuộc tính mà cả $\boldsymbol{x}$ và $\boldsymbol{y}$ đều có, và $\|\boldsymbol{x}\|$ và $\|\boldsymbol{y}\|$ lần lượt là căn bậc hai (*trung bình hình học*) của số lượng thuộc tính mà $\boldsymbol{x}$ và $\boldsymbol{y}$ có. Do đó, sim $(\boldsymbol{x}, \boldsymbol{y})$ là thước đo tương đối về mức độ sở hữu các thuộc tính chung.

Một biến thể đơn giản của độ tương đồng cosine trong trường hợp trên là:

$$\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\boldsymbol{x} \cdot \boldsymbol{x} + \boldsymbol{y} \cdot \boldsymbol{y} - \boldsymbol{x} \cdot \boldsymbol{y}}, \tag{2.29}$$

đây là tỷ lệ giữa số lượng thuộc tính chung mà $\boldsymbol{x}$ và $\boldsymbol{y}$ có và số lượng thuộc tính mà

x hoặc ***y*** có. Hàm này, được gọi là **hệ số Tanimoto** hay **khoảng cách Tanimoto**, thường được sử dụng trong truy xuất thông tin và phân loại sinh học.

2.3.8 Độ tương đồng giữa các phân phối: độ phân kỳ Kullback – Leibler

Cuối cùng, chúng ta giới thiệu *độ phân kỳ Kullback – Leibler* (*độ phân kỳ KL*), một thước đo được sử dụng phổ biến trong khai phá dữ liệu để đo lường sự khác biệt giữa hai phân phối xác suất trên cùng một biến x. Khái niệm này có nguồn gốc từ lý thuyết xác suất và lý thuyết thông tin.

Độ phân kỳ KL, liên quan chặt chẽ đến khái niệm *entropy tương đối*, *độ phân kỳ thông tin*, và *thông tin phân biệt*, là một thước đo phi đối xứng của sự khác biệt giữa hai phân phối xác suất $p(x)$ và $q(x)$. Cụ thể, độ phân kỳ KL của $q(x)$ so với $p(x)$, được ký hiệu là $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big)$, là thước đo mức độ mất mát thông tin khi sử dụng $q(x)$ để xấp xỉ $p(x)$.

Giả sử $p(x)$ và $q(x)$ là hai phân phối xác suất của một biến ngẫu nhiên rời rạc x. Điều này có nghĩa là cả $p(x)$ và $q(x)$ đều có tổng bằng 1, và với mọi x thuộc tập X, ta có $p(x) > 0$ và $q(x) > 0$. $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big)$ được định nghĩa theo công thức:

$$D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}. \tag{2.30}$$

Độ phân kỳ KL đo lường giá trị kỳ vọng (giá trị trung bình) của số lượng bit dư thừa cần thiết để mã hóa các mẫu từ $p(x)$ khi sử dụng một bộ mã hóa dựa trên $q(x)$ thay vì sử dụng bộ mã hóa dựa trên $p(x)$. Thông thường, $p(x)$ biểu diễn phân phối "thực" của dữ liệu, các quan sát, hoặc một phân phối lý thuyết tính toán chính xác. Thước đo $q(x)$ thường biểu diễn một lý thuyết, mô hình, mô tả, hoặc sự xấp xỉ của $p(x)$.

Phiên bản liên tục của độ phân kỳ KL có dạng:

$$D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx. \tag{2.31}$$

Mặc dù độ phân kỳ KL đo lường "khoảng cách" giữa hai phân phối, nhưng nó không phải là một thước đo khoảng cách theo nghĩa metric, vì: nó không đối xứng: $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big)$ thường không bằng $D_{\mathrm{KL}}\big(q(x) \parallel p(x)\big)$, và không thỏa mãn bất đẳng thức tam giác. Tuy nhiên, $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big)$ luôn không âm, và $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big) = 0$ nếu và chỉ nếu $p(x) = q(x)$.

Lưu ý rằng cần cẩn trọng khi tính toán độ phân kỳ KL. Chúng ta biết rằng $\lim_{p(x)\to 0} p(x)\log p(x) = 0$. Tuy nhiên, nếu $p(x) \neq 0$ mà $q(x) = 0$, thì $D_{\mathrm{KL}}\big(p(x) \parallel q(x)\big)$ được định nghĩa là ∞. Điều này có nghĩa là nếu một biến cố e có khả năng xảy ra (tức $p(e) > 0$) nhưng $q(e) = 0$ (tức $q(e)$ cho rằng e là tuyệt đối không thể xảy ra), thì hai phân phối là hoàn toàn khác nhau. Tuy nhiên, trong thực tế, hai phân phối P và Q được suy ra từ các quan sát và đếm mẫu, tức từ các phân phối tần suất. Do đó, sẽ không hợp lý khi dự đoán trong phân phối xác suất thu được rằng một biến cố là hoàn toàn không thể xảy ra, vì chúng ta phải tính đến khả năng xuất hiện các biến cố chưa được thấy. Một phương pháp *làm mịn* có thể được sử dụng để suy ra phân phối xác suất từ phân phối tần suất quan sát được, như được minh họa trong ví dụ sau.

Ví dụ 30 (Tính độ phân kỳ KL bằng phương pháp làm mịn)**.** Giả sử có hai phân phối mẫu P và Q như sau: $P : \left(a : \frac{3}{5},\ b : \frac{1}{5},\ c : \frac{1}{5}\right)$ và $Q : \left(a : \frac{5}{9},\ b : \frac{3}{9},\ d : \frac{1}{9}\right)$. Để tính độ phân kỳ $D_{\mathrm{KL}}\big(P \parallel Q\big)$, ta đưa ra một hằng số nhỏ ε, ví dụ $\varepsilon = 10^{-3}$, và định nghĩa phiên bản làm mịn của P và Q, gọi là P' và Q' như sau.

Tập hợp các ký hiệu quan sát được trong P là $S_P = \{a, b, c\}$. Tương tự, tập hợp các ký hiệu quan sát được trong Q là $S_Q = \{a, b, d\}$. Hợp của hai tập trên là $S_U = \{a, b, c, d\}$. Bằng cách làm mịn, các ký hiệu bị thiếu có thể được thêm vào mỗi phân phối với xác suất nhỏ ε. Do đó, ta có: $P' : \left(a : \frac{3}{5} - \frac{\varepsilon}{3},\ b : \frac{1}{5} - \frac{\varepsilon}{3},\ c : \frac{1}{5} - \frac{\varepsilon}{3},\ d : \varepsilon\right)$ và $Q' : \left(a : \frac{5}{9} - \frac{\varepsilon}{3},\ b : \frac{3}{9} - \frac{\varepsilon}{3},\ c : \varepsilon,\ d : \frac{1}{9} - \frac{\varepsilon}{3}\right)$. Từ đó, dễ dàng tính được $D_{\mathrm{KL}}\big(P' \parallel Q'\big)$ theo công thức (2.30).

2.3.9 Nắm bắt ý nghĩa ẩn trong các thước đo tương đồng

Thước đo tương đồng là một khái niệm cơ bản trong khai phá dữ liệu. Chúng ta đã giới thiệu nhiều thước đo để tính mức độ tương đồng giữa các đối tượng, bao gồm các thuộc tính số, thuộc tính nhị phân đối xứng và bất đối xứng, thuộc tính thứ tự và thuộc tính danh định. Chúng ta cũng đã giới thiệu cách tính tương đồng giữa các tài liệu sử dụng mô hình không gian vector, cũng như cách so sánh hai phân phối bằng khái niệm độ phân kỳ KL. Những khái niệm và thước đo về sự tương

đồng của đối tượng này sẽ được sử dụng rất nhiều trong các nghiên cứu tiếp theo về các phương pháp khám phá mẫu, phân loại, phân cụm và phân tích điểm ngoại lai.

Trong các ứng dụng thực tế, ta có thể gặp khái niệm về sự tương đồng của đối tượng vượt ra ngoài những gì đã được bàn ở chương này. Ngay cả đối với các đối tượng đơn giản, sự tương đồng giữa các đối tượng thường liên quan chặt chẽ đến ý nghĩa ngữ nghĩa của chúng, điều này không thể được nắm bắt chỉ dựa trên các thước đo tương đồng đã định nghĩa ở trên. Ví dụ, người ta thường cho rằng hình học và đại số có sự tương đồng cao hơn so với hình học với âm nhạc hay chính trị, mặc dù tất cả đều là các môn học được giảng dạy ở trường. Hơn nữa, các tài liệu có phân phối tần suất từ (hoặc "bag of words") tương tự nhau có thể thể hiện những ý nghĩa khá khác nhau (ví dụ, xét "The cat bites a mouse" so với "The mouse bites a cat"). Điều này vượt ra ngoài khả năng xử lý của mô hình không gian vector đã được trình bày ở Mục 2.3.7.

Hơn nữa, các đối tượng có thể được cấu thành từ các cấu trúc và mối quan hệ phức tạp. Các thước đo tương đồng dành cho đồ thị và mạng lưới có thể cần được giới thiệu, điều này vượt ra ngoài những khái niệm về sự tương đồng của đối tượng đã được giới thiệu ở đây.

Trong các chương sắp tới, chúng ta sẽ giới thiệu thêm các thước đo tương đồng khi gặp phải trong các bài toán và phương pháp được bàn luận. Đặc biệt, ở Chương 12, chúng ta sẽ giới thiệu ngắn gọn về khái niệm biểu diễn phân phối và học biểu diễn (representation learning), trong đó việc nhúng văn bản (text embedding) và học sâu sẽ được sử dụng để tính toán những khái niệm tương đồng nâng cao như vậy.

2.4 Chất lượng dữ liệu, làm sạch dữ liệu và tích hợp dữ liệu

2.4.1 Thước đo chất lượng dữ liệu

2.4.2 Làm sạch dữ liệu

2.4.3 Tích hợp dữ liệu

2.5 Biến đổi dữ liệu

2.5.1 Chuẩn hóa

2.5.2 Rời rạc hóa

2.5.3 Nén dữ liệu

2.5.4 Lấy mẫu

2.6 Giảm chiều dữ liệu

2.6.1 Phân tích thành phần chính

2.6.2 Lựa chọn tập con thuộc tính

2.6.3 Các phương pháp giảm chiều phi tuyến

2.7 Tóm tắt

2.8 Bài tập

9.

2.9 Tài liệu tham khảo