các phương pháp phân cụm. Phần thảo luận về các phương pháp phân cụm nâng cao được dành cho Chương 11.

# Phân tích cluster

Phần này thiết lập cơ sở để nghiên cứu phân tích cụm. Phần 10.1.1 xác định phân tích cụm và trình bày các ví dụ về nơi nó hữu ích. Trong Phần 10.1.2, bạn sẽ tìm hiểu các khía cạnh để so sánh các phương pháp phân cụm, cũng như các yêu cầu đối với phân cụm. Tổng quan về các kỹ thuật phân cụm cơ bản được trình bày trong Phần 10.1.3.

## Phân tích cụm là gì?

**Phân tích cluster** hay đơn giản là phân cụm là quá trình phân vùng một tập hợp các đối tượng dữ liệu (hoặc các quan sát) thành các tập con. Mỗi tập hợp con là một cụm, sao cho các đối tượng trong một cụm tương tự với nhau, nhưng không giống với các đối tượng trong các cụm khác. Tập hợp các cụm kết quả từ một phân tích cụm có thể được gọi là phân cụm. Trong bối cảnh này, các phương pháp phân cụm khác nhau có thể tạo ra các nhóm khác nhau trên cùng một tập dữ liệu. Việc phân vùng không phải do con người thực hiện mà do thuật toán phân cụm. Do đó, phân nhóm hữu ích ở chỗ nó có thể dẫn đến việc phát hiện ra các nhóm chưa biết trước đó trong dữ liệu.

Phân tích cụm đã được sử dụng rộng rãi trong nhiều ứng dụng như hiểu biết kinh doanh, nhận dạng mẫu hình ảnh, tìm kiếm trên web, sinh học và bảo mật. Trong kinh doanh thông minh, phân nhóm có thể được sử dụng để tổ chức một số lượng lớn khách hàng thành các nhóm, nơi các khách hàng trong một nhóm có những đặc điểm giống nhau mạnh mẽ. Điều này tạo điều kiện thuận lợi cho việc phát triển các chiến lược kinh doanh để tăng cường quản lý quan hệ khách hàng. Hơn nữa, hãy xem xét một công ty tư vấn với một số lượng lớn các dự án. Để cải thiện quản lý dự án, phân cụm có thể được áp dụng để phân chia các dự án thành các danh mục dựa trên sự tương đồng để việc đánh giá và chẩn đoán dự án (để cải thiện việc phân phối và kết quả dự án) có thể được tiến hành một cách hiệu quả.

Trong nhận dạng hình ảnh, phân cụm có thể được sử dụng để khám phá các cụm hoặc "Lớp con" trong hệ thống nhận dạng ký tự viết tay. Giả sử chúng ta có một tập dữ liệu gồm các chữ số viết tay, trong đó mỗi chữ số được gắn nhãn là 1, 2, 3, v.v. Lưu ý rằng có thể có một phương sai lớn trong cách mọi người viết cùng một chữ số. Lấy ví dụ số 2. Một số người có thể viết nó bằng một vòng tròn nhỏ ở phần dưới cùng bên trái, trong khi một số người khác thì không. Chúng ta có thể sử dụng phân cụm để xác định các lớp con cho “2”, mỗi lớp đại diện cho một biến thể về cách mà 2 có thể được viết. Sử dụng nhiều mô hình dựa trên các lớp con có thể cải thiện độ chính xác nhận dạng tổng thể.

Clustering cũng đã tìm thấy nhiều ứng dụng trong tìm kiếm Web. Ví dụ: một tìm kiếm từ khóa thường có thể trả về một số lượng truy cập rất lớn (tức là các trang có liên quan đến tìm kiếm) do số lượng trang web cực kỳ lớn. Clustering có thể được sử dụng để tổ chức các kết quả tìm kiếm thành các nhóm và trình bày kết quả một cách ngắn gọn và dễ tiếp cận. Hơn nữa, các kỹ thuật phân cụm đã được phát triển để phân cụm tài liệu thành các chủ đề, được sử dụng phổ biến trong thực hành truy xuất thông tin.

Là một chức năng khai thác dữ liệu, phân tích cụm có thể được sử dụng như một công cụ độc lập để hiểu rõ hơn về việc phân phối dữ liệu, quan sát các đặc điểm của từng cụm và tập trung vào một nhóm cụ thể để phân tích thêm. Ngoài ra, nó có thể phục vụ như một bước tiền xử lý cho các thuật toán khác, chẳng hạn như mô tả đặc tính, lựa chọn tập hợp con thuộc tính và phân loại, sau đó sẽ hoạt động trên các cụm được phát hiện và các thuộc tính hoặc tính năng đã chọn.

Bởi vì một cụm là một tập hợp các đối tượng dữ liệu tương tự với nhau trong cụm và khác với các đối tượng trong các cụm khác, một cụm đối tượng dữ liệu có thể được coi như một lớp ngầm định. Theo nghĩa này, phân cụm đôi khi được gọi là phân nhóm tự động. Một lần nữa, một điểm khác biệt quan trọng ở đây là việc phân cụm có thể tự động tìm các nhóm. Đây là một lợi thế khác biệt của phân tích cụm.

Phân cụm còn được gọi là phân đoạn dữ liệu trong một số ứng dụng vì phân cụm phân vùng các tập dữ liệu lớn thành các nhóm theo sự giống nhau của chúng. Phân cụm cũng có thể là đã sử dụng for **ngoại lệ dò tìm**, wđây ngoại lai (giá trị cái đó Chúng tôi “xa ađường" từ bất kỳ cụm nào) có thể thú vị hơn các trường hợp thông thường. Các ứng dụng phát hiện ngoại lệ bao gồm phát hiện gian lận thẻ tín dụng và giám sát các hoạt động tội phạm trong thương mại điện tử. Ví dụ, các trường hợp ngoại lệ trong giao dịch thẻ tín dụng, chẳng hạn như mua hàng rất đắt và không thường xuyên, có thể được quan tâm là các hoạt động gian lận có thể xảy ra. Phát hiện ngoại lệ là chủ đề của Chương 12.

Phân cụm dữ liệu đang được phát triển mạnh mẽ. Các lĩnh vực nghiên cứu đóng góp bao gồm khai thác dữ liệu, thống kê, học máy, công nghệ cơ sở dữ liệu không gian, truy xuất thông tin, tìm kiếm trên web, sinh học, tiếp thị và nhiều lĩnh vực ứng dụng khác. Do lượng dữ liệu khổng lồ được thu thập trong cơ sở dữ liệu, phân tích cụm gần đây đã trở thành một chủ đề tích cực trong nghiên cứu khai thác dữ liệu.

Là một nhánh của thống kê, phân tích cụm đã được nghiên cứu rộng rãi, với trọng tâm chính là phân tích cụm dựa trên khoảng cách. Các công cụ phân tích cụm dựa trên k-means, k-medoids và một số phương pháp khác cũng đã được tích hợp vào nhiều gói hoặc hệ thống phần mềm phân tích thống kê, chẳng hạn như S-Plus, SPSS và SAS. Trong học máy, hãy nhớ lại rằng phân loại được gọi là học có giám sát vì thông tin nhãn lớp được đưa ra, nghĩa là thuật toán học được giám sát trong đó nó được thông báo cho thành viên lớp của mỗi bộ huấn luyện. Phân cụm được gọi là học không giám sát vì thông tin nhãn lớp không có. Vì lý do này, phân cụm là một hình thức học bằng cách quan sát, thay vì học bằng ví dụ. Trong khai thác dữ liệu, các nỗ lực đã tập trung vào việc tìm kiếm các phương pháp phân tích cụm hiệu quả và hiệu quả trong cơ sở dữ liệu lớn. Các chủ đề nghiên cứu tích cực tập trung vào khả năng mở rộng của các phương pháp phân cụm, tính hiệu quả của các phương pháp phân nhóm các hình dạng phức tạp (ví dụ: không lồi) và các loại dữ liệu (ví dụ: văn bản, đồ thị và hình ảnh), kỹ thuật phân nhóm nhiều chiều (ví dụ: phân cụm đối tượng với hàng nghìn tính năng), và các phương pháp phân nhóm dữ liệu số và dữ liệu danh nghĩa hỗn hợp trong cơ sở dữ liệu lớn.

## Yêu cầu đối với phân tích cụm

Phân cụm là một lĩnh vực nghiên cứu đầy thách thức. Trong phần này, bạn sẽ tìm hiểu về các yêu cầu đối với phân cụm như một công cụ khai thác dữ liệu, cũng như các khía cạnh có thể được sử dụng để so sánh các phương pháp phân cụm.

Sau đây là các yêu cầu điển hình của phân cụm trong khai thác dữ liệu.

**Khả năng mở rộng**: Nhiều thuật toán phân cụm hoạt động tốt trên các tập dữ liệu nhỏ chứa ít hơn vài trăm đối tượng dữ liệu; tuy nhiên, một cơ sở dữ liệu lớn có thể chứa hàng triệu hoặc thậm chí hàng tỷ đối tượng, đặc biệt là trong các kịch bản tìm kiếm trên Web. Chỉ tập hợp trên một mẫu của tập dữ liệu lớn nhất định có thể dẫn đến kết quả sai lệch. Do đó, các thuật toán phân cụm có khả năng mở rộng cao là cần thiết.

**Khả năng đối phó với các loại thuộc tính khác nhau**: Nhiều thuật toán được thiết kế để phân cụm dữ liệu số (dựa trên khoảng thời gian). Tuy nhiên, các ứng dụng có thể yêu cầu nhóm các kiểu dữ liệu khác, chẳng hạn như dữ liệu nhị phân, danh nghĩa (phân loại) và thứ tự hoặc hỗn hợp của các kiểu dữ liệu này. Gần đây, ngày càng có nhiều ứng dụng cần đến kỹ thuật phân cụm cho các kiểu dữ liệu phức tạp như đồ thị, chuỗi, hình ảnh và tài liệu.

**Khám phá các cụm có hình dạng tùy ý**: Nhiều thuật toán phân cụm xác định các cụm dựa trên các phép đo khoảng cách Euclidean hoặc Manhattan (Chương 2). Các thuật toán dựa trên các phép đo khoảng cách như vậy có xu hướng tìm các cụm hình cầu có kích thước và mật độ tương tự. Tuy nhiên, một cụm có thể có bất kỳ hình dạng nào. Hãy xem xét các cảm biến, ví dụ, thường được triển khai để giám sát môi trường. Phân tích cụm trên các bài đọc của cảm biến có thể phát hiện ra các hiện tượng thú vị. Chúng ta có thể muốn sử dụng phân cụm để tìm ranh giới của một đám cháy rừng đang diễn ra, thường không phải là hình cầu. Điều quan trọng là phát triển các thuật toán có thể phát hiện các cụm có hình dạng tùy ý.

**Yêu cầu về kiến ​​thức miền để xác định các tham số đầu vào**: Nhiều thuật toán phân cụm yêu cầu người dùng cung cấp kiến ​​thức miền dưới dạng các tham số đầu vào như số lượng cụm mong muốn. Do đó, kết quả phân cụm có thể nhạy cảm với các tham số như vậy. Thông số thường khó xác định, đặc biệt là đối với các tập dữ liệu có chiều cao và nơi người dùng chưa hiểu sâu về dữ liệu của họ. Việc yêu cầu đặc tả kiến ​​thức miền không chỉ tạo gánh nặng cho người dùng mà còn khiến chất lượng phân cụm khó kiểm soát.

**Khả năng đối phó với dữ liệu nhiễu**: Hầu hết các tập dữ liệu trong thế giới thực đều chứa các dữ liệu ngoại lai và / hoặc bị thiếu, không xác định hoặc sai. Ví dụ, các kết quả đọc của cảm biến thường bị nhiễu - một số kết quả đọc có thể không chính xác do cơ chế cảm biến và một số kết quả đọc có thể bị sai do nhiễu từ các vật thể thoáng qua xung quanh. Thuật toán phân cụm có thể nhạy cảm với nhiễu như vậy và có thể tạo ra các cụm chất lượng kém. Do đó, chúng ta cần các phương pháp phân cụm mạnh mẽ với nhiễu.

**Phân cụm gia tăng và không nhạy cảm với thứ tự đầu vào**: Trong nhiều ứng dụng, các bản cập nhật gia tăng (đại diện cho dữ liệu mới hơn) có thể đến bất kỳ lúc nào. Một số thuật toán phân cụm không thể kết hợp các cập nhật gia tăng vào cấu trúc phân nhóm hiện có và thay vào đó, phải tính toán lại một phân nhóm mới từ đầu. Các thuật toán phân cụm cũng có thể nhạy cảm với thứ tự dữ liệu đầu vào. Có nghĩa là, với một tập hợp các đối tượng dữ liệu, các thuật toán phân cụm có thể trả về các chuỗi khác nhau đáng kể tùy thuộc vào thứ tự mà các đối tượng được trình bày. Các thuật toán phân cụm tăng dần và các thuật toán không nhạy cảm với thứ tự đầu vào là cần thiết.

**Khả năng phân nhóm dữ liệu chiều cao**: Một tập dữ liệu có thể chứa nhiều thứ nguyên hoặc thuộc tính. Ví dụ: khi phân nhóm tài liệu, mỗi từ khóa có thể được coi là một thứ nguyên và thường có hàng nghìn từ khóa. Hầu hết các thuật toán phân cụm đều tốt trong việc xử lý dữ liệu chiều thấp, chẳng hạn như tập dữ liệu chỉ liên quan đến hai hoặc ba chiều. Việc tìm kiếm các cụm đối tượng dữ liệu trong một không gian chiều cao là một thách thức, đặc biệt là khi xem xét rằng dữ liệu đó có thể rất thưa thớt và rất sai lệch.

**Phân cụm dựa trên ràng buộc**: Các ứng dụng trong thế giới thực có thể cần thực hiện phân nhóm dưới nhiều loại ràng buộc khác nhau. Giả sử rằng công việc của bạn là chọn vị trí cho một số máy rút tiền tự động (ATM) mới nhất định trong một thành phố. Để quyết định điều này, bạn có thể phân cụm các hộ gia đình trong khi xem xét các ràng buộc như hệ thống sông ngòi và đường cao tốc của thành phố cũng như loại và số lượng khách hàng trên mỗi cụm. Một nhiệm vụ đầy thách thức là tìm các nhóm dữ liệu có hành vi phân cụm tốt đáp ứng các ràng buộc đã chỉ định.

**Khả năng diễn giải và khả năng sử dụng**: Người dùng muốn kết quả phân nhóm có thể diễn giải, dễ hiểu và có thể sử dụng được. Nghĩa là, phân cụm có thể cần phải gắn liền với các ứng dụng và diễn giải ngữ nghĩa cụ thể. Điều quan trọng là phải nghiên cứu mục tiêu ứng dụng có thể ảnh hưởng như thế nào đến việc lựa chọn các tính năng phân cụm và phương pháp phân nhóm.

Sau đây là các khía cạnh trực giao mà các phương pháp phân cụm có thể được so sánh với:

**Tiêu chí phân vùng**: Trong một số phương thức, tất cả các đối tượng được phân vùng để không tồn tại hệ thống phân cấp giữa các cụm. Có nghĩa là, tất cả các cụm đều ở cùng một mức độ về mặt khái niệm. Một phương pháp như vậy rất hữu ích, chẳng hạn, để phân chia khách hàng thành các nhóm để mỗi nhóm có người quản lý riêng. Ngoài ra, các phương pháp khác phân chia các đối tượng dữ liệu một cách phân cấp, trong đó các cụm có thể được hình thành ở các mức ngữ nghĩa khác nhau. Ví dụ: trong khai thác văn bản, chúng tôi có thể muốn tổ chức một kho tài liệu thành multiple general đếnbức ảnh, Such như “chính trị" và “các môn thể thao," mỗi of which có thể có chủ đề phụ, For ví dụ, “Football, ” “bóng rổ, ” “bóng chày," và “khúc côn cầu" có thể exist làm chủ đề phụ of “các môn thể thao." Các sau này fcủa chúng tôi Subtopics Chúng tôi tại a thấp hơn cấp độ trong các phân cấp hơn “các môn thể thao."

**Tách các cụm**: Một số phương pháp phân vùng các đối tượng dữ liệu thành các cụm loại trừ lẫn nhau. Khi gom khách hàng thành từng nhóm để mỗi nhóm do một người quản lý chăm sóc, mỗi khách hàng chỉ được thuộc một nhóm. Trong một số trường hợp khác, các cụm có thể không độc quyền, nghĩa là một đối tượng dữ liệu có thể thuộc về nhiều hơn một cụm. Ví dụ, khi nhóm các tài liệu thành các chủ đề, một tài liệu có thể liên quan đến nhiều chủ đề. Do đó, các chủ đề dưới dạng cụm có thể không độc quyền.

**Thước đo tương tự**: Một số phương pháp xác định độ giống nhau giữa hai đối tượng bằng khoảng cách giữa chúng. Khoảng cách như vậy có thể được xác định trên không gian Euclide,

mạng lưới đường, không gian vectơ hoặc bất kỳ không gian nào khác. Trong các phương pháp khác, sự giống nhau có thể được xác định bằng kết nối dựa trên mật độ hoặc sự tiếp giáp, và có thể không dựa trên khoảng cách tuyệt đối giữa hai đối tượng. Các biện pháp tương tự đóng một vai trò cơ bản trong việc thiết kế các phương pháp phân cụm. Trong khi các phương pháp dựa trên khoảng cách thường có thể tận dụng các kỹ thuật tối ưu hóa, các phương pháp dựa trên mật độ và tính liên tục thường có thể tìm thấy các cụm có hình dạng tùy ý.

**Phân cụm không gian**: Nhiều phương pháp phân cụm tìm kiếm các cụm trong toàn bộ không gian dữ liệu nhất định. Các phương pháp này rất hữu ích cho các tập dữ liệu có kích thước thấp. Tuy nhiên, với dữ liệu chiều cao, có thể có nhiều thuộc tính không liên quan, có thể làm cho các phép đo tương tự không đáng tin cậy. Do đó, các cụm được tìm thấy trong không gian đầy đủ là oftenvô nghĩa. Thay vào đó, tốt hơn là nên tìm kiếm các cụm trong các không gian con khác nhau của cùng một tập dữ liệu. Phân cụm không gian con phát hiện ra các cụm và không gian con (thường có số chiều thấp) biểu hiện sự giống nhau của đối tượng.

Đến kết luận, thuật toán phân cụm có một số yêu cầu. Những yếu tố này bao gồm khả năng mở rộng và khả năng đối phó với các loại thuộc tính khác nhau, dữ liệu nhiễu, cập nhật gia tăng, các cụm có hình dạng tùy ý và các ràng buộc. Khả năng diễn giải và khả năng sử dụng cũng rất quan trọng. Ngoài ra, các phương pháp phân cụm có thể khác nhau đối với cấp độ phân vùng, cho dù các cụm có loại trừ lẫn nhau hay không, các biện pháp tương tự được sử dụng và việc phân cụm không gian con có được thực hiện hay không.

## Tổng quan về các phương pháp phân cụm cơ bản

Có rất nhiều thuật toán phân cụm trong tài liệu. Rất khó để cung cấp một phân loại rõ ràng về các phương pháp phân cụm vì các danh mục này có thể chồng chéo lên nhau do đó một phương pháp có thể có các tính năng từ một số danh mục. Tuy nhiên, sẽ hữu ích khi trình bày một bức tranh tương đối có tổ chức về các phương pháp phân cụm. Nói chung, các phương pháp phân cụm cơ bản chính có thể được phân loại thành các loại sau, được thảo luận trong phần còn lại của chương này.

**Các phương pháp phân vùng:** Cho một tập hợp n đối tượng, một phương pháp phân vùng xây dựng k phân vùng dữ liệu, trong đó mỗi phân vùng đại diện cho một cụm và k n. Tức là nó chia dữ liệu thành k nhóm sao cho mỗi nhóm phải chứa ít nhất một đối tượng. Nói cách khác, các phương pháp phân vùng tiến hành phân vùng một cấp trên các tập dữ liệu. Các phương pháp phân vùng cơ bản thường áp dụng phân tách cụm độc quyền. Tức là mỗi đối tượng phải thuộc đúng một nhóm. Yêu cầu này có thể được nới lỏng, ví dụ, trong kỹ thuật phân vùng mờ. Tham chiếu đến các kỹ thuật này được nêu trong ghi chú thư mục (Phần 10.9).

≤

Hầu hết các phương pháp phân vùng đều dựa trên khoảng cách. Cho k, số phân vùng cần tạo, một phương pháp phân vùng sẽ tạo ra một phân vùng ban đầu. Sau đó, nó sử dụng kỹ thuật tái định vị lặp đi lặp lại để cố gắng cải thiện việc phân vùng bằng cách di chuyển các đối tượng từ nhóm này sang nhóm khác. Tiêu chí chung của một phân vùng tốt là ovật bắn trong các tương tự cluster Chúng tôi “Cthua" or rphấn khởi to mỗi khác, wđây là oném vào khác nhau clusters Chúng tôi “xa riêng biệt" or very khác nhau. Đó Chúng tôi đa dạng các loại of khác

tiêu chí đánh giá chất lượng của vách ngăn. Các phương pháp phân vùng truyền thống có thể được mở rộng để phân cụm không gian con, thay vì tìm kiếm toàn bộ không gian dữ liệu. Điều này rất hữu ích khi có nhiều thuộc tính và dữ liệu thưa thớt.

Để đạt được tính tối ưu toàn cục trong phân nhóm dựa trên phân vùng thường bị cấm tính toán, có khả năng đòi hỏi phải liệt kê đầy đủ tất cả các phân vùng có thể có. Thay vào đó, hầu hết các ứng dụng đều áp dụng các phương pháp heuristic phổ biến, chẳng hạn như các phương pháp tiếp cận tham lam như k-means và k-medoids, giúp cải thiện chất lượng phân cụm và tiếp cận tối ưu cục bộ. Các phương pháp phân cụm heuristic này hoạt động hiệu quả để tìm kiếm các cụm hình cầu trong cơ sở dữ liệu có kích thước vừa và nhỏ. Để tìm các cụm có hình dạng phức tạp và cho các tập dữ liệu rất lớn, cần mở rộng phương pháp dựa trên phân vùng. Các phương pháp phân cụm dựa trên phân vùng được nghiên cứu sâu trong Phần 10.2.

**Các phương pháp phân cấp:** Một phương pháp phân cấp tạo ra sự phân rã có thứ bậc của tập các đối tượng dữ liệu đã cho. Một phương pháp phân cấp có thể được phân loại là tích hợp hoặc phân chia, dựa trên cách hình thành sự phân rã thứ bậc. Cách tiếp cận tích tụ, còn được gọi là cách tiếp cận từ dưới lên, bắt đầu với mỗi đối tượng tạo thành một nhóm riêng biệt. Nó liên tiếp hợp nhất các đối tượng hoặc nhóm gần nhau, cho đến khi tất cả các nhóm được hợp nhất thành một (cấp cao nhất của hệ thống phân cấp) hoặc điều kiện kết thúc được giữ nguyên. Phương pháp phân chia, còn được gọi là phương pháp từ trên xuống, bắt đầu với tất cả các đối tượng trong cùng một cụm. Trong mỗi lần lặp lại liên tiếp, một cụm được chia thành các cụm nhỏ hơn, cho đến khi cuối cùng mỗi đối tượng nằm trong một cụm hoặc điều kiện kết thúc được giữ nguyên.

Các phương pháp phân cụm phân cấp có thể dựa trên khoảng cách hoặc dựa trên mật độ và tính liên tục. Các phần mở rộng khác nhau của các phương pháp phân cấp cũng xem xét việc phân cụm trong không gian con.

Các phương thức phân cấp có thực tế là một khi một bước (hợp nhất hoặc tách) được thực hiện, nó không bao giờ có thể được hoàn tác. Độ cứng này hữu ích ở chỗ nó dẫn đến chi phí so sánh nhỏ hơn do không phải lo lắng về số lượng tổ hợp các lựa chọn khác nhau. Những kỹ thuật như vậy không thể sửa chữa những quyết định sai lầm; tuy nhiên, các phương pháp để cải thiện chất lượng của phân cụm phân cấp đã được đề xuất. Các phương pháp phân cụm theo thứ bậc được nghiên cứu trong Phần 10.3.

**Phương pháp dựa trên mật độ:** Hầu hết các phương pháp phân vùng tập hợp các đối tượng dựa trên sự khác biệt giữa các đối tượng. Các phương pháp như vậy chỉ có thể tìm thấy các cụm hình cầu và gặp khó khăn trong việc phát hiện các cụm có hình dạng tùy ý. Các phương pháp phân cụm khác đã được phát triển dựa trên khái niệm về mật độ. Ý tưởng chung của họ là tiếp tục phát triển một cụm nhất định miễn là mật độ (số lượng đối tượng hoặc dữ liệu điểm) trong các “khu vực lân cận" exceeds một số ngưỡng cửa. For example, for mỗi điểm dữ liệu trong một cụm nhất định, vùng lân cận của một bán kính nhất định phải chứa ít nhất một số điểm tối thiểu. Phương pháp như vậy có thể được sử dụng để lọc ra nhiễu hoặc các giá trị ngoại lai và phát hiện các cụm có hình dạng tùy ý.

Các phương pháp dựa trên mật độ có thể chia một tập hợp các đối tượng thành nhiều nhóm riêng biệt, hoặc một hệ thống phân cấp của các cụm. Thông thường, các phương pháp dựa trên mật độ chỉ xem xét các cụm độc quyền và không xem xét các cụm mờ. Hơn nữa, các phương pháp dựa trên mật độ có thể được mở rộng từ không gian đầy đủ sang phân cụm không gian con. Các phương pháp phân cụm dựa trên mật độ được nghiên cứu trong Phần 10.4.

**Các phương pháp dựa trên lưới:** Phương pháp dựa trên lưới lượng tử hóa không gian đối tượng thành một số lượng hữu hạn ô tạo thành cấu trúc lưới. Tất cả các hoạt động phân cụm đều được hình thành trên cấu trúc lưới (tức là trên không gian lượng tử hóa). Ưu điểm chính của phương pháp này là thời gian xử lý nhanh, thường không phụ thuộc vào số lượng đối tượng dữ liệu và chỉ phụ thuộc vào số lượng ô trong mỗi chiều trong không gian lượng tử hóa.

Sử dụng lưới thường là một cách tiếp cận hiệu quả cho nhiều vấn đề khai thác dữ liệu không gian, bao gồm phân cụm. Do đó, các phương pháp dựa trên lưới có thể được tích hợp với các phương pháp phân cụm khác như phương pháp dựa trên mật độ và phương pháp phân cấp. Phân cụm dựa trên lưới được nghiên cứu trong Phần 10.5.

Các phương pháp này được tóm tắt ngắn gọn trong Hình 10.1. Một số thuật toán phân cụm tích hợp các ý tưởng của một số phương pháp phân cụm, do đó, đôi khi rất khó để xác định một thuật toán đã cho là duy nhất thuộc về một loại phương pháp phân nhóm. Hơn nữa, một số ứng dụng có thể có các tiêu chí phân cụm yêu cầu tích hợp một số kỹ thuật phân cụm.

Trong các phần sau, chúng ta sẽ xem xét chi tiết từng phương pháp phân cụm. Các phương pháp phân cụm nâng cao và các vấn đề liên quan được thảo luận trong Chương 11. Nói chung, ký hiệu được sử dụng như sau. Gọi D là tập dữ liệu gồm n đối tượng được phân cụm. Một đối tượng được mô tả bởi d biến, trong đó mỗi biến còn được gọi là thuộc tính hoặc thứ nguyên,

|  |  |
| --- | --- |
| **phương pháp** | **Đặc điểm chung** |
| Các phương pháp phân vùng | * Tìm các cụm hình cầu loại trừ lẫn nhau * Dựa trên khoảng cách * có thể sử dụng trung bình hoặc trung bình (v.v.) để đại diện cho trung tâm cụm * Hiệu quả đối với các tập dữ liệu có kích thước vừa và nhỏ |
| Phương pháp phân cấp | * Clustering là một phân tách theo thứ bậc (tức là nhiều cấp) * Không thể sửa lỗi hợp nhất hoặc tách sai * có thể kết hợp các kỹ thuật khác như microclustering hoặc   consider object “liên kết ” |
| Phương pháp dựa trên mật độ | * Có thể tìm thấy các cụm có hình dạng tùy ý * Các cụm là các vùng dày đặc của các đối tượng trong không gian được ngăn cách bởi các vùng có mật độ thấp * Mật độ cụm: Mỗi điểm phải có số điểm tối thiểu within nó là “khu vực lân cận" * có thể lọc ra các ngoại lệ |
| Các phương pháp dựa trên lưới | * Sử dụng một cấu trúc dữ liệu lưới đa phân giải * Thời gian xử lý nhanh (thường không phụ thuộc vào số lượng   đối tượng dữ liệu, nhưng phụ thuộc vào kích thước lưới) |

**Hình 10.1** Tổng quan về các phương pháp phân cụm được thảo luận trong chương này. Lưu ý rằng một số thuật toán có thể kết hợp nhiều phương pháp khác nhau.

và do đó cũng có thể được gọi là một điểm trong không gian vật thể d chiều. Các đối tượng được thể hiện bằng phông chữ in nghiêng đậm (ví dụ: p).

# Phương pháp phân vùng

Phiên bản đơn giản nhất và cơ bản nhất của phân tích cụm là phân vùng, tổ chức các đối tượng của một tập hợp thành một số nhóm hoặc cụm độc quyền. Để giữ cho đặc tả vấn đề ngắn gọn, chúng ta có thể giả định rằng số lượng cụm được đưa ra như là kiến ​​thức nền tảng. Tham số này là điểm bắt đầu cho các phương pháp phân vùng.

Về mặt hình thức, cho một tập dữ liệu, D, gồm n đối tượng và k, số lượng cụm sẽ tạo thành, một thuật toán phân vùng sắp xếp các đối tượng thành k phân vùng (kn), trong đó mỗi phân vùng đại diện cho một cụm. Các cụm được hình thành để tối ưu hóa tiêu chí phân vùng khách quan, chẳng hạn như hàm khác biệt dựa trên khoảng cách, để các đối tượng trong một cụm "tương tự" với nhau và "không giống" với các đối tượng trong các cụm khác về thuộc tính tập dữ liệu .

≤

Trong phần này, bạn sẽ tìm hiểu các phương pháp phân vùng nổi tiếng và thường được sử dụng nhất — k-means (Phần 10.2.1) và k-medoid (Phần 10.2.2). Bạn cũng sẽ tìm hiểu một số biến thể của các phương pháp phân vùng cổ điển này và cách chúng có thể được mở rộng để xử lý các tập dữ liệu lớn.

## *k*-Phương tiện: Một kỹ thuật dựa trên Centroid

Giả sử một tập dữ liệu, D, chứa n đối tượng trong không gian Euclide. Các phương pháp phân vùng phân chia các đối tượng trong D thành k cụm, C1,. . . , Ck, tức là Ci D và Ci Cj cho (1 i, jk). Một hàm mục tiêu được sử dụng để đánh giá chất lượng phân vùng sao cho các đối tượng trong một cụm tương tự với nhau nhưng khác với các đối tượng trong các cụm khác. Đây là, hàm mục tiêu nhằm mục đích cho độ giống nhau giữa các lớp cao và độ tương tự giữa các lớp thấp.

≤ ≤

⊂∩ = ∅

Kỹ thuật phân vùng dựa trên centroid sử dụng tâm của một cụm, Ci, để đại diện cho cụm đó. Về mặt khái niệm, tâm của một cụm là điểm trung tâm của nó. Trọng tâm có thể được xác định theo nhiều cách khác nhau như trung bình hoặc trung bình của các đối tượng (hoặc điểm) được gán cho cụm. Sự khác biệt giữa một đối tượng p ∈ Ci và ci, đại diện của cụm, được đo bằng dist (p, ci), trong đó dist (x, y) là khoảng cách Euclide giữa hai điểm x và y. Chất lượng của cụm Ci có thể được đo lường bằng sự biến thiên bên trong cụm, là tổng sai số bình phương giữa tất cả các đối tượng trong Ci và ci centroid, được định nghĩa là

*k*

Σ Σ

*E* = *quận(* ***p***, ci) 2, (10.1)

*Tôi*= 1 *p*∈*CTôi*

trong đó E là tổng sai số bình phương cho tất cả các đối tượng trong tập dữ liệu; p là điểm trong không gian biểu diễn một đối tượng nhất định; và ci là tâm của cụm Ci (cả p và ci đều đa chiều). Nói cách khác, đối với mỗi đối tượng trong mỗi cụm, khoảng cách từ

đối tượng đến trung tâm cụm của nó được bình phương và các khoảng cách được tính tổng. Hàm mục tiêu này cố gắng làm cho k cụm thu được càng nhỏ gọn và càng tách biệt càng tốt.

Việc tối ưu hóa biến thể trong cụm là một thách thức về mặt tính toán. Trong trường hợp xấu nhất, chúng tôi sẽ phải liệt kê một số phân vùng có thể có tính lũy thừa cho số cụm và kiểm tra các giá trị biến thể trong cụm. Nó đã được chỉ ra rằng vấn đề là NP-khó trong không gian Euclid nói chung, ngay cả đối với hai cụm (tức là, k 2). Hơn nữa, vấn đề là khó NP đối với một số cụm k tổng quát ngay cả trong không gian Euclide 2-D. Nếu số cụm k và số chiều của không gian d là cố định, bài toán có thể được giải quyết trong thời gian O (ndk + 1 log n), với n là số đối tượng. Để khắc phục chi phí tính toán quá lớn cho giải pháp chính xác, các phương pháp tiếp cận tham lam thường được sử dụng trong thực tế. Một ví dụ chính là thuật toán k-mean, thuật toán này đơn giản và thường được sử dụng.

=

*“Làm thế nào để* thuật toán k-mean hoạt động? " Thuật toán k-mean xác định tâm của một cụm là giá trị trung bình của các điểm trong cụm. Nó tiến hành như sau. Đầu tiên, nó chọn ngẫu nhiên k trong số các đối tượng trong D, mỗi đối tượng ban đầu đại diện cho một trung tâm hoặc trung tâm cụm. Đối với mỗi đối tượng còn lại, một đối tượng được gán cho cụm mà nó giống nhất, dựa trên khoảng cách Euclide giữa đối tượng và giá trị trung bình của cụm. Sau đó, thuật toán k-mean cải thiện lặp đi lặp lại các biến thể trong cụm. Đối với mỗi cụm, nó tính toán giá trị trung bình mới bằng cách sử dụng các đối tượng được gán cho cụm trong lần lặp trước. Tất cả các đối tượng sau đó được chỉ định lại bằng cách sử dụng các phương tiện được cập nhật làm trung tâm cụm mới. Các lần lặp lại tiếp tục cho đến khi việc gán ổn định, nghĩa là các cụm được hình thành trong vòng hiện tại cũng giống như các cụm được hình thành trong vòng trước. Quy trình k-mean được tóm tắt trong Hình 10.2.

**Thuật toán: *k***-có nghĩa.Các thuật toán k-mean để phân vùng, trong đó trung tâm của mỗi cụm được biểu diễn bằng giá trị trung bình của các đối tượng trong cụm.

**Đầu vào:**

*k*: số lượng cụm,

*D*: một tập dữ liệu chứa n đối tượng.

**Đầu ra:** Một tập hợp gồm k cụm.

**Phương pháp:**

1. tùy ý chọn k đối tượng từ D làm tâm cụm ban đầu;
2. **nói lại**
3. (lại) gán từng đối tượng vào cụm mà đối tượng tương tự nhất, dựa trên giá trị trung bình của các đối tượng trong cụm;
4. cập nhật các phương tiện của cụm, tức là tính giá trị trung bình của các đối tượng cho mỗi cụm;
5. **cho đến khi** không thay đổi;

**Hình 10.2** Thuật toán phân vùng k-mean.



















* 1. Phân cụm ban đầu **(b)** Lặp lại **(c)** Nhóm cuối cùng

**Hình 10.3** Phân cụm tập hợp các đối tượng bằng phương pháp k-means; cho (b) cập nhật các trung tâm cụm và gán lại các đối tượng cho phù hợp (giá trị trung bình của mỗi cụm được đánh dấu bằng dấu +).

**Ví dụ 10.1 Phân cụm theo *k*-means phân vùng.** Hãy xem xét một tập hợp các đối tượng nằm trong không gian 2-D, như được mô tả trong Hình 10.3 (a). Giả sử k 3, nghĩa là người dùng muốn các đối tượng được phân chia thành ba cụm.

=

Theo thuật toán trong Hình 10.2, chúng ta tùy ý chọn ba đối tượng làm ba tâm cụm ban đầu, trong đó các tâm cụm được đánh dấu bằng a. Mỗi đối tượng được gán cho một cụm dựa trên trung tâm cụm mà nó là đối tượng gần nhất. Sự phân bố như vậy tạo thành các bóng được bao quanh bởi các đường cong chấm chấm, như trong Hình 10.3 (a).

+

Tiếp theo, các trung tâm cụm được cập nhật. Nghĩa là, giá trị trung bình của mỗi cụm được tính toán lại dựa trên các đối tượng hiện tại trong cụm. Sử dụng các trung tâm cụm mới, các đối tượng được phân phối lại cho các cụm dựa trên trung tâm cụm nào là gần nhất. Sự phân bố lại như vậy tạo thành các bóng mới được bao quanh bởi các đường cong đứt nét, như trong Hình 10.3 (b).

Quá trình này lặp lại, dẫn đến Hình 10.3 (c). Quá trình phân bổ lại lặp đi lặp lại các đối tượng thành các cụm để cải thiện việc phân vùng được gọi là tái định vị lặp lại. Đồng đều, không có sự gán lại các đối tượng trong bất kỳ cụm nào xảy ra và do đó quá trình kết thúc. Các cụm kết quả được trả về bởi quá trình phân cụm.

Phương thức k-mean không được đảm bảo hội tụ đến mức tối ưu toàn cục và thường kết thúc ở mức tối ưu cục bộ. Kết quả có thể phụ thuộc vào sự lựa chọn ngẫu nhiên ban đầu của các trung tâm cụm. (Bạn sẽ được yêu cầu đưa ra một ví dụ để thể hiện điều này như một bài tập.) Để có được kết quả tốt trong thực tế, người ta thường chạy thuật toán k-mean nhiều lần với các trung tâm cụm ban đầu khác nhau.

Độ phức tạp theo thời gian của thuật toán k-mean là O (nkt), trong đó n là tổng số đối tượng, k là số cụm và t là số lần lặp. Thông thường, kn và t n. Do đó, phương pháp này có khả năng mở rộng tương đối và hiệu quả trong việc xử lý các tập dữ liệu lớn.

Có một số biến thể của phương pháp k-mean. Những điều này có thể khác nhau trong việc lựa chọn phương tiện k ban đầu, tính toán độ không đồng đều và các chiến lược để tính toán phương tiện cụm.

Phương pháp k-mean chỉ có thể được áp dụng khi giá trị trung bình của một tập đối tượng được xác định. Điều này có thể không đúng trong một số ứng dụng chẳng hạn như khi dữ liệu có thuộc tính danh nghĩa có liên quan. Phương pháp k-mode là một biến thể của k-means, mở rộng mô hình k-mean thành dữ liệu danh nghĩa của cụm bằng cách thay thế các phương tiện của cụm bằng các chế độ. Nó sử dụng các biện pháp khác biệt mới để xử lý các đối tượng danh nghĩa và một phương pháp dựa trên tần số để cập nhật các chế độ của các cụm. Các phương pháp k-means và k-mode có thể được tích hợp để phân cụm dữ liệu với các giá trị danh nghĩa và số hỗn hợp.

Sự cần thiết cho người dùng để chỉ định k, số lượng cụm, trước có thể được xem là một bất lợi. Tuy nhiên, đã có những nghiên cứu về cách khắc phục khó khăn này, chẳng hạn như bằng cách cung cấp khoảng giá trị k gần đúng, và sau đó sử dụng kỹ thuật phân tích để xác định k tốt nhất bằng cách so sánh kết quả phân cụm thu được với các giá trị k khác nhau. Phương pháp k-mean không thích hợp để phát hiện các cụm có hình dạng không lồi hoặc các cụm có kích thước rất khác nhau. Hơn nữa, nó nhạy cảm với nhiễu và các điểm dữ liệu ngoại lệ vì một số lượng nhỏ dữ liệu như vậy có thể ảnh hưởng đáng kể đến giá trị trung bình. "Làm thế nào chúng tôi có thể làm cho thuật toán k-mean có thể mở rộng hơn?" Một cách tiếp cận để khai thác phương pháp k-means hiệu quả hơn trên các tập dữ liệu lớn là sử dụng một tập mẫu có kích thước tốt trong phân nhóm. Một phương pháp khác là sử dụng một phương pháp lọc sử dụng chỉ mục dữ liệu lưu trữ cấp độ không gian để tiết kiệm chi phí khi tính toán. Phương pháp tiếp cận thứ ba khám phá sự tích tụ vi ý tưởng, which Đầu tiên groups gần đây ovật bắn thành “vi cụm ” và sau đó thực hiện phân cụm k-mean trên các vi cụm. Microclustering được thảo luận thêm

trong Mục 10.3.

## *k*-Medoids: Một kỹ thuật dựa trên đối tượng đại diện

Thuật toán k-mean nhạy cảm với các giá trị ngoại lệ vì các đối tượng như vậy nằm xa phần lớn dữ liệu và do đó, khi được gán cho một cụm, chúng có thể làm sai lệch đáng kể giá trị trung bình của cụm. Điều này vô tình ảnh hưởng đến việc gán các đối tượng khác cho các cụm. Hiệu ứng này đặc biệt trầm trọng hơn do việc sử dụng hàm sai số bình phương của Eq. (10.1), như được quan sát trong Ví dụ 10.2.

**Ví dụ 10.2 Một nhược điểm của *k*-có nghĩa.** Xét sáu điểm trong không gian 1-D có các giá trị tương ứng là 1, 2, 3, 8, 9, 10 và 25. Bằng trực giác, bằng cách kiểm tra trực quan, chúng ta có thể hình dung

các điểm được phân chia thành các cụm 1, 2, 3 và 8, 9, 10, trong đó điểm 25 bị loại trừ vì nó có vẻ là một điểm ngoại lệ. K-mean sẽ phân vùng các giá trị như thế nào? Nếu chúng ta áp dụng k-mean bằng cách sử dụng k 2 và Eq. (10.1), phân vùng 1, 2, 3, 8, 9, 10, 25 có biến thể trong cụm

{} { }

= {{} { }}

*(*1 - 2) 2 + (2 - 2) 2 + (3 - 2) 2 + (8 - 13) 2 + (9 - 13) 2 + (10 - 13) 2 + (25 - 13) 2 = 196,

cho rằng trung bình của cụm 1, 2, 3 là 2 và trung bình của 8, 9, 10, 25 là 13. So sánh

{} { }

điều này cho phân vùng 1, 2, 3, 8, 9, 10, 25, mà k-mean tính toán biến thể trong cụm như

{{} { }}

*(*1 - 3,5) 2 + (2 - 3,5) 2 + (3 - 3,5) 2 + (8 - 3,5) 2 + (9 - 14,67) 2

+ *(*10 - 14,67) + (25 - 14,67) = 189,67,

2

2

cho rằng 3,5 là giá trị trung bình của cụm 1, 2, 3, 8 và 14,67 là giá trị trung bình của cụm 9, 10, 25. Phân vùng thứ hai có sự thay đổi trong cụm thấp nhất; do đó, phương thức k-mean gán giá trị 8 cho một cụm khác với cụm chứa 9 và 10 do điểm ngoại lệ 25. Hơn nữa, tâm của cụm thứ hai, 14,67, về cơ bản là xa tất cả các thành viên trong cụm.

{} { }

*"Làm thế nào chúng ta có thể sửa đổi thuật toán k-mean để giảm độ nhạy như vậy đối với các ngoại lệ?"* Thay vì lấy giá trị trung bình của các đối tượng trong một cụm làm điểm tham chiếu, chúng ta có thể chọn các đối tượng thực tế để đại diện cho các cụm, sử dụng một đối tượng đại diện cho mỗi cụm. Mỗi đối tượng còn lại được gán cho cụm mà đối tượng đại diện là giống nhất. Sau đó, phương pháp phân vùng được thực hiện dựa trên nguyên tắc tối thiểu hóa tổng các điểm khác biệt giữa mỗi đối tượng p và đối tượng đại diện tương ứng của nó. Đó là, tiêu chí lỗi tuyệt đối được sử dụng, được định nghĩa là

*k*

Σ Σ

*E* = *quận(****p***, oi), (10,2)

*Tôi*= 1 ***p***∈*CTôi*

trong đó E là tổng của sai số tuyệt đối cho tất cả các đối tượng p trong tập dữ liệu, và oi là đối tượng đại diện của Ci. Đây là cơ sở cho phương pháp k-medoids, nhóm n đối tượng thành k cụm bằng cách giảm thiểu sai số tuyệt đối (Phương trình 10.2).

Khi k 1, chúng ta có thể tìm được trung vị chính xác trong thời gian O (n2). Tuy nhiên, khi k là một

=

số dương tổng quát, bài toán k-medoid là NP-khó.

Thuật toán Phân vùng xung quanh Medoids (PAM) (xem Hình 10.5 sau) là một hiện thực hóa đơn giản của việc phân cụm k-medoid. Nó giải quyết vấn đề theo cách lặp đi lặp lại, tham lam. Giống như thuật toán k-mean, các đối tượng đại diện ban đầu (được gọi là hạt giống) được chọn tùy ý. Chúng tôi xem xét liệu việc thay thế một đối tượng đại diện bằng một đối tượng không phản hồi có cải thiện chất lượng phân nhóm hay không. Tất cả các thay thế có thể được thử. Quá trình lặp đi lặp lại của việc thay thế các đối tượng đại diện bởi các đối tượng khác tiếp tục cho đến khi chất lượng của việc phân cụm kết quả không thể được cải thiện bởi bất kỳ sự thay thế nào. Chất lượng này được đo bằng một hàm chi phí của sự khác biệt trung bình giữa một đối tượng và đối tượng đại diện trong cụm của nó.

Cụ thể, hãy để o1 ,. . . , ok là tập hợp các đối tượng đại diện hiện tại (ví dụ: medoid). Để xác định xem một đối tượng không đại diện, được ký hiệu là orandom, có phải là vật thay thế tốt cho oj trung bình hiện tại (1 ≤ j ≤ k) hay không, chúng tôi tính khoảng cách từ mọi đối tượng p đến đối tượng gần nhất trong tập hợp o1 ,. . . , oj − 1, orandom, oj + 1 ,. . . , ok, và sử dụng khoảng cách để cập nhật hàm chi phí. Việc gán lại các đối tượng cho

{ }

{***o*1**, *. . .*, oj − 1, orandom, oj + 1, *. . .*, ok} rất đơn giản. Giả sử đối tượng p hiện đang được gán cho một cụm được đại diện bởi medoid oj (Hình 10.4a hoặc b). Chúng ta có cần gán lại p cho một cụm khác không nếu oj được thay thế bằng orandom? Đối tượng p cần được gán lại cho hoặc hoặc một số cụm khác được đại diện bởi oi (i/ = *j)*, cái nào là gần nhất. Ví dụ, trong Hình 10.4 (a), p gần nhất với oi và do đó được gán lại cho oi. Tuy nhiên, trong Hình 10.4 (b), p gần với orandom nhất và do đó được gán lại cho orandom. Điều gì xảy ra nếu thay vào đó, p hiện được gán cho một cụm được đại diện bởi một số đối tượng khác oi, tôi/ = *j*?

* + - 1. Đã giao lại cho o***Tôi***

***ongẫu nhiên***

***j***

***o***

***p***

***oTôi***

* + - 1. Đã giao lại quá***ngẫu nhiên***

***oTôi***

***oj***

***p***

***ongẫu nhiên***

* + - 1. Không thay đổi

***oTôi***

***oj***

***p***

***ongẫu nhiên***

* + - 1. Đã giao lại quá***ngẫu nhiên***

Đối tượng dữ liệu Trung tâm cụm trước khi hoán đổi

Sau khi hoán đổi

***ongẫu nhiên***

***p***

***oj***

***oTôi***

**Hình 10.4** Bốn trường hợp của hàm chi phí cho nhóm k-medoids.

Đối tượng o vẫn được gán cho cụm được đại diện bởi oi miễn là o vẫn gần oi hơn là orandom (Hình 10.4c). Nếu không, o được gán lại thành orandom (Hình 10.4d).

Mỗi lần chuyển nhượng lại xảy ra, một sự khác biệt về sai số tuyệt đối, E, được đóng góp vào hàm chi phí. Do đó, hàm chi phí tính toán sự khác biệt về giá trị sai số tuyệt đối nếu một đối tượng đại diện hiện tại được thay thế bằng một đối tượng không đại diện. Tổng chi phí hoán đổi là tổng chi phí phát sinh của tất cả các đối tượng không mang tính đại diện. Nếu tổng chi phí là số âm, thì oj được thay thế hoặc hoán đổi bằng orandom vì sai số tuyệt đối thực tế E giảm xuống. Nếu tổng chi phí là dương, đối tượng đại diện hiện tại, oj, được coi là có thể chấp nhận được và không có gì thay đổi trong lần lặp.

*“Phương pháp nào hiệu quả hơn — k-mean hay k-medoids?”* Phương pháp k-medoids mạnh mẽ hơn k-mean khi có nhiễu và giá trị ngoại lệ vì medoid ít bị ảnh hưởng bởi các giá trị ngoại lệ hoặc các giá trị cực trị khác hơn là giá trị trung bình. Tuy nhiên, độ phức tạp của mỗi lần lặp trong thuật toán k-medoids là O (k (nk) 2). Đối với các giá trị lớn của n và k, việc tính toán như vậy trở nên rất tốn kém và tốn kém hơn nhiều so với phương pháp k-mean. Cả hai phương pháp đều yêu cầu người dùng chỉ định k, số lượng cụm.

-

*"Làm thế nào chúng ta có thể mở rộng quy mô của phương pháp k-medoids?"* Thuật toán phân vùng k-medoids điển hình như PAM (Hình 10.5) hoạt động hiệu quả cho các tập dữ liệu nhỏ, nhưng không chia tỷ lệ tốt cho các tập dữ liệu lớn. Để xử lý các tập dữ liệu lớn hơn, một phương pháp dựa trên lấy mẫu được gọi là CLARA (Clustering LARge Applications) có thể được sử dụng. Thay vì xem xét toàn bộ tập dữ liệu, CLARA sử dụng một mẫu ngẫu nhiên của tập dữ liệu. Thuật toán PAM sau đó được áp dụng để tính toán các medoid tốt nhất từ ​​mẫu. Lý tưởng nhất là mẫu phải đại diện chặt chẽ cho tập dữ liệu gốc. Trong nhiều trường hợp, một mẫu lớn hoạt động tốt nếu nó được tạo ra để mỗi đối tượng có xác suất được chọn vào mẫu bằng nhau. Các đối tượng đại diện (medoid) được chọn sẽ có khả năng tương tự như những đối tượng đã được chọn từ toàn bộ tập dữ liệu. CLARA xây dựng các nhóm từ nhiều mẫu ngẫu nhiên và trả về nhóm tốt nhất làm đầu ra. Độ phức tạp của việc tính toán medoid trên một mẫu ngẫu nhiên là O (ks2 k (nk)), trong đó s là kích thước của mẫu, k là số cụm và n là tổng số đối tượng. CLARA có thể xử lý các tập dữ liệu lớn hơn PAM.

+ -

Hiệu quả của CLARA phụ thuộc vào cỡ mẫu. Lưu ý rằng PAM tìm kiếm k-medoid tốt nhất trong số một tập dữ liệu nhất định, trong khi CLARA tìm kiếm k-medoid tốt nhất trong số mẫu được chọn của tập dữ liệu. CLARA không thể tìm thấy một nhóm tốt nếu bất kỳ loại thuốc nào tốt nhất được lấy mẫu khác xa loại thuốc k-medoid tốt nhất. Nếu một đối tượng

**Thuật toán:** *k***-mô hình.** PAM, một thuật toán k-medoids để phân vùng dựa trên medoid hoặc các đối tượng trung tâm.

**Đầu vào:**

*k*: số lượng cụm,

*D*: một tập dữ liệu chứa n đối tượng.

**Đầu ra:** Một tập hợp gồm k cụm.

**Phương pháp:**

* + - * 1. tự ý chọn k vật ở D làm vật hoặc hạt đại diện ban đầu;
        2. **nói lại**
        3. gán từng đối tượng còn lại vào cụm có đối tượng đại diện gần nhất;
        4. chọn ngẫu nhiên một đối tượng không đại diện, o***ngẫu nhiên***;
        5. tính tổng chi phí, S, của việc hoán đổi đối tượng đại diện, o***j***, với o***ngẫu nhiên***;
        6. **nếu** *S <* 0 rồi hoán đổi o***j*** với o***ngẫu nhiên*** để hình thành tập k đối tượng đại diện mới;
        7. **cho đến khi** không thay đổi;

**Hình 10.5** PAM, một thuật toán phân vùng k-medoids.

là một trong những loại k-medoid tốt nhất nhưng không được chọn trong quá trình lấy mẫu, CLARA sẽ không bao giờ tìm ra nhóm tốt nhất. (Bạn sẽ được yêu cầu cung cấp một ví dụ minh họa điều này như một bài tập.)

*“Làm cách nào chúng ta có thể cải thiện chất lượng và khả năng mở rộng của CLARA?”* Nhớ lại rằng khi tìm kiếm các medoid tốt hơn, PAM kiểm tra mọi đối tượng trong tập dữ liệu so với mọi medoid hiện tại, trong khi CLARA giới hạn các medoid ứng viên chỉ là một mẫu ngẫu nhiên của tập dữ liệu. Một thuật toán ngẫu nhiên được gọi là CLARANS (Phân cụm các ứng dụng lớn dựa trên Tìm kiếm được phân nhóm theo RAN) trình bày sự cân bằng giữa chi phí và hiệu quả của việc sử dụng các mẫu để có được phân cụm.

Đầu tiên, nó chọn ngẫu nhiên k đối tượng trong tập dữ liệu làm medoid hiện tại. Sau đó, nó chọn ngẫu nhiên một medoid hiện tại x và một đối tượng y không phải là một trong những medoid hiện tại. Thay x bằng y có thể cải thiện tiêu chí lỗi tuyệt đối không? Nếu có, thay thế sẽ được thực hiện. CLARANS thực hiện một cuộc tìm kiếm ngẫu nhiên như vậy l lần. Tập hợp các medoid hiện tại sau bước l được coi là tối ưu cục bộ. CLARANS lặp lại quá trình ngẫu nhiên này m lần và trả về kết quả cuối cùng là giá trị tối ưu cục bộ tốt nhất.

# Phương pháp phân cấp

Trong khi các phương pháp phân vùng đáp ứng yêu cầu phân nhóm cơ bản là tổ chức một tập hợp các đối tượng thành một số nhóm độc quyền, trong một số trường hợp, chúng ta có thể muốn phân vùng dữ liệu của mình thành các nhóm ở các cấp độ khác nhau, chẳng hạn như trong một hệ thống phân cấp. Phương pháp phân cụm phân cấploks by grouping dữ liệu ovật bắn thành a hệ thống cấp bậc or “Tree ” of clchùm.

Việc biểu diễn các đối tượng dữ liệu dưới dạng một hệ thống phân cấp rất hữu ích cho việc tóm tắt và trực quan hóa dữ liệu. Ví dụ: với tư cách là giám đốc nhân sự tại AllElectronics,

bạn có thể sắp xếp nhân viên của mình thành các nhóm chính như giám đốc điều hành, người quản lý và nhân viên. Bạn có thể phân vùng thêm các nhóm này thành các nhóm con nhỏ hơn. Ví dụ, nhóm nhân viên chung có thể được chia thành các nhóm nhỏ gồm các sĩ quan cấp cao, sĩ quan và học viên. Tất cả các nhóm này tạo thành một hệ thống phân cấp. Chúng ta có thể dễ dàng tóm tắt hoặc mô tả đặc điểm của dữ liệu được sắp xếp thành một hệ thống phân cấp, có thể được sử dụng để tìm, chẳng hạn như mức lương trung bình của các nhà quản lý và của viên chức.

Hãy xem xét nhận dạng ký tự viết tay như một ví dụ khác. Đầu tiên, một tập hợp các mẫu viết tay có thể được phân chia thành các nhóm chung trong đó mỗi nhóm tương ứng với một ký tự duy nhất. Một số nhóm có thể được phân chia thêm thành các nhóm con vì một ký tự có thể được viết theo nhiều cách khác nhau về cơ bản. Nếu cần thiết, phân vùng phân cấp có thể được tiếp tục đệ quy cho đến khi đạt được độ chi tiết mong muốn.

Trong các ví dụ trước, mặc dù chúng tôi đã phân vùng dữ liệu theo thứ bậc, nhưng chúng tôi không giả định rằng dữ liệu có cấu trúc phân cấp (ví dụ: người quản lý ở cùng cấp trong hệ thống phân cấp AllElectronics của chúng tôi với tư cách là nhân viên). Việc chúng tôi sử dụng hệ thống phân cấp ở đây chỉ là để tóm tắt và biểu diễn dữ liệu cơ bản theo cách nén. Hệ thống phân cấp như vậy đặc biệt hữu ích cho việc trực quan hóa dữ liệu.

Ngoài ra, trong một số ứng dụng, chúng tôi có thể tin rằng dữ liệu có cấu trúc phân cấp khó hiểu mà chúng tôi muốn khám phá. Ví dụ: phân cụm phân cấp có thể khám phá ra một hệ thống phân cấp cho các nhân viên của AllElectronics được cấu trúc dựa trên tiền lương. Trong nghiên cứu về sự tiến hóa, phân nhóm theo thứ bậc có thể phân nhóm các loài động vật theo các đặc điểm logic sinh học của chúng để khám phá ra các con đường tiến hóa, đó là một hệ thống phân cấp của các loài. Một ví dụ khác, việc nhóm các cấu hình của trò chơi chiến lược (ví dụ: cờ vua hoặc cờ caro) theo cách phân cấp có thể giúp phát triển các chiến lược trò chơi có thể được sử dụng để huấn luyện người chơi. Trong phần này, bạn sẽ nghiên cứu các phương pháp phân cụm phân cấp. Phần 10.3.1 bắt đầu bằng cuộc thảo luận về phân nhóm phân cấp so với phân nhóm, tổ chức các đối tượng thành một hệ thống phân cấp bằng cách sử dụng chiến lược từ dưới lên hoặc từ trên xuống, tương ứng. Các phương thức tổng hợp bắt đầu với các đối tượng riêng lẻ dưới dạng các cụm, được hợp nhất lặp đi lặp lại để tạo thành các cụm lớn hơn. Ngược lại, các phương thức chia ban đầu cho phép tất cả các đối tượng đã cho

tạo thành một cụm, chúng được chia lặp đi lặp lại thành các cụm nhỏ hơn.

Các phương pháp phân cụm phân cấp có thể gặp khó khăn liên quan đến việc lựa chọn các điểm hợp nhất hoặc tách. Một quyết định như vậy là rất quan trọng, bởi vì khi một nhóm đối tượng được hợp nhất hoặc tách ra, quy trình ở bước tiếp theo sẽ hoạt động trên các cụm mới được tạo. Nó sẽ không hoàn tác những gì đã làm trước đó, cũng không thực hiện hoán đổi đối tượng giữa các cụm. Do đó, các quyết định hợp nhất hoặc chia tách, nếu không được lựa chọn tốt, có thể dẫn đến các cụm chất lượng thấp. Hơn nữa, các phương pháp này không chia tỷ lệ tốt vì mỗi quyết định hợp nhất hoặc chia tách cần phải kiểm tra và đánh giá nhiều đối tượng hoặc cụm.

Một hướng đầy hứa hẹn để cải thiện chất lượng phân nhóm của các phương thức phân cấp là tích hợp phân nhóm phân cấp với các kỹ thuật phân nhóm khác, dẫn đến **nhiều pha** (hoặc nhiều pha) phân cụm. Chúng tôi giới thiệu hai phương pháp như vậy, đó là BIRCH và Chameleon. BIRCH (Phần 10.3.3) bắt đầu bằng cách phân vùng các đối tượng theo thứ bậc bằng cách sử dụng cấu trúc cây, trong đó các nút lá hoặc nút không lá ở mức thấp có thể được xem như “các vi cụm” tùy thuộc vào thang độ phân giải. Sau đó, nó áp dụng khác

các thuật toán phân cụm để thực hiện phân nhóm macro trên các vi cụm. Chameleon (Phần 10.3.4) khám phá mô hình động trong phân cụm phân cấp.

Có một số cách trực giao để phân loại các phương pháp phân cụm phân cấp. Ví dụ, chúng có thể được phân loại thành các phương pháp thuật toán, phương pháp xác suất và phương pháp Bayes. Các phương pháp tổng hợp, chia chia và đa pha là thuật toán, có nghĩa là chúng coi các đối tượng dữ liệu là các cụm xác định và tính toán theo khoảng cách xác định giữa các đối tượng. Phương pháp xác suất sử dụng các mô hình xác suất để nắm bắt các cụm và đo lường chất lượng của các cụm bằng sự phù hợp của các mô-đun. Chúng ta thảo luận về phân cụm phân cấp theo xác suất trong Phần 10.3.5. Các phương pháp Bayes tính toán một phân phối của các cụm có thể có. Nghĩa là, thay vì xuất ra một cụm xác định duy nhất trên một tập dữ liệu, chúng trả về một nhóm cấu trúc phân cụm và xác suất của chúng, có điều kiện dựa trên dữ liệu đã cho.

## Phân cụm phân chia theo thứ bậc và tổng hợp

Phương pháp phân nhóm phân cấp có thể là cộng gộp hoặc phân chia, tùy thuộc vào việc phân rã phân cấp được hình thành theo kiểu từ dưới lên (hợp nhất) hay từ trên xuống (tách) thời trang. Let's có a ckẻ thất bại nhìn tại những các chiến lược.

Phương pháp phân nhóm phân cấp tích tụ sử dụng chiến lược từ dưới lên. Thông thường, nó bắt đầu bằng cách để mỗi đối tượng tạo thành cụm riêng của nó và kết hợp lặp đi lặp lại các cụm thành các cụm lớn hơn và lớn hơn, cho đến khi tất cả các đối tượng nằm trong một cụm duy nhất hoặc một số điều kiện nhất định được thỏa mãn. Cụm đơn trở thành gốc của hệ thống phân cấp. Đối với bước hợp nhất, nó tìm hai cụm gần nhau nhất (theo một số biện pháp tương tự) và kết hợp cả hai để tạo thành một cụm. Bởi vì hai cụm được hợp nhất cho mỗi lần lặp, trong đó mỗi cụm chứa ít nhất một đối tượng, một phương thức kết hợp yêu cầu tối đa n lần lặp.

Phương pháp phân nhóm phân cấp sử dụng chiến lược từ trên xuống. Nó bắt đầu bằng cách đặt tất cả các đối tượng vào một cụm, là gốc của hệ thống phân cấp. Sau đó, nó chia cụm gốc thành một số cụm con nhỏ hơn và phân vùng đệ quy các cụm đó thành những cụm nhỏ hơn. Quá trình phân vùng tiếp tục cho đến khi mỗi cụm ở cấp thấp nhất đủ chặt chẽ — hoặc chỉ chứa một đối tượng hoặc các đối tượng trong một cụm đủ giống nhau.

Trong phân nhóm phân cấp hoặc phân nhóm, người dùng có thể chỉ định số lượng cụm mong muốn làm điều kiện kết thúc.

**Ví dụ 10.3 Phân nhóm theo thứ bậc tổng hợp so với phân chia.** Hình 10.6 cho thấy ứng dụng của AGNES (AGglomerative NESting), một phương pháp phân nhóm phân cấp tích tụ và DIANA (DIvisive ANAlysis), một phương pháp phân nhóm phân cấp có chia hết, trên tập dữ liệu gồm năm đối tượng, a, b, c, d, e . Ban đầu, AGNES, phương pháp tích tụ, đặt mỗi đối tượng vào một cụm của riêng nó. Các cụm sau đó được hợp nhất từng bước theo một số tiêu chí. Ví dụ: các cụm C1 và C2 có thể được hợp nhất nếu một đối tượng ở C1 và một đối tượng ở C2 tạo thành khoảng cách Euclide tối thiểu giữa hai đối tượng bất kỳ từ

{ }