

10 Phân tích cụm: Cơ bản Khái niệm và phương pháp

Hãy tưởng tượng rằng bạn là Giám đốc Quan hệ Khách hàng tại AllElectronics và bạn có năm người quản lý làm việc cho mình. Bạn muốn sắp xếp tất cả khách hàng của công ty thành năm nhóm để mỗi nhóm có thể được giao cho một người quản lý khác nhau. Về mặt chiến lược, bạn muốn khách hàng trong mỗi nhóm càng giống nhau càng tốt. Hơn nữa, hai khách hàng nhất định có mô hình kinh doanh rất khác nhau không nên được xếp vào cùng một nhóm. Ý định của bạn dường như chiến lược kinh doanh này là phát triển các dịch vụ quan hệ khách hàng nhằm mục tiêu cụ thể vào từng nhóm, dựa trên các đặc điểm chung mà khách hàng trong mỗi nhóm chia sẻ. Những loại kỹ thuật khai thác dữ liệu nào có thể giúp bạn hoàn thành nhiệm vụ này?

Không giống như trong phân loại, nhãn lớp (hoặc ID nhóm) của mỗi khách hàng là không xác định. Bạn cần khám phá những nhóm này. Với số lượng lớn khách hàng và nhiều thuộc tính mô tả hồ sơ khách hàng, việc để con người nghiên cứu dữ liệu và đưa ra cách phân chia khách hàng thành các nhóm chiến lược theo cách thủ công có thể rất tốn kém hoặc thậm chí không khả thi. Bạn cần một công cụ phân cụm để hỗ trợ.

Phân cụm là quá trình nhóm một tập hợp các đối tượng dữ liệu thành nhiều nhóm hoặc cụm sao cho các đối tượng trong một cụm có độ tương đồng cao, nhưng rất khác biệt với các đối tượng trong các cụm khác. Sự khác biệt và tương đồng được đánh giá dựa trên giá trị thuộc tính sử dụng để mô tả các đối tượng và thường liên quan đến các phép đo khoảng cách.¹ Phân cụm như một công cụ khai thác dữ liệu có nguồn gốc từ nhiều lĩnh vực ứng dụng như sinh học, bảo mật, trí tuệ kinh doanh và tìm kiếm trên Web.

Chương này trình bày các khái niệm và phương pháp cơ bản của phân tích cụm. Trong Phần 10.1, chúng tôi giới thiệu chủ đề và nghiên cứu các yêu cầu của phương pháp phân cụm cho lượng dữ liệu lớn và nhiều ứng dụng khác nhau. Bạn sẽ học một số kỹ thuật phân cụm cơ bản, được tổ chức thành các loại sau: phương pháp phân vùng (Phần 10.2), phương pháp phân cấp (Phần 10.3), phương pháp dựa trên mật độ (Phần 10.4) và phương pháp dựa trên lưới (Phần 10.5). Trong Phần 10.6, chúng tôi thảo luận ngắn gọn về cách đánh giá

¹Sự tương đồng và khác biệt của dữ liệu được thảo luận chi tiết trong Phần 2.4. Bạn có thể muốn tham khảo phần đó để xem nhanh.

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

10.1 Phân tích cụm

Phần này thiết lập nền tảng để nghiên cứu phân tích cụm. Phần 10.1.1 định nghĩa phân tích cụm và trình bày các ví dụ về nơi phân tích cụm 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 cụ cơ bản được trình bày trong Phần 10.1.3.

10.1.1 Phân tích cụm là gì?

Phân tích cụm hay đơn giản là phân cụm là quá trình phân vùng một tập hợp các đối tượng dữ liệu (hoặc quan sát) thành các tập hợp con. Mỗi tập hợp con là một cụm, sao cho các đối tượng trong một cụm tương tự nhau, nhưng không giống với các đối tượng trong các cụm khác. Tập hợp các cụm kết quả từ 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 cụm khác nhau trên cùng một tập dữ liệu. Phân vùng không được thực hiện bởi con người, mà bởi thuật toán phân cụm. Do đó, phân cụm hữu ích ở chỗ nó có thể dẫn đến việc khám phá ra các nhóm trước đây chưa biết trong dữ liệu.

Phân tích cụm đã được sử dụng rộng rãi trong nhiều ứng dụng như trí tuệ kinh doanh, nhận dạng mẫu hình ảnh, tìm kiếm trên web, sinh học và bảo mật. Trong trí tuệ kinh doanh, phân cụm có thể được sử dụng để sắp xếp một số lượng lớn khách hàng thành các nhóm, trong đó khách hàng trong một nhóm có chung các đặc điểm tương đồng mạnh mẽ. Điều này tạo điều kiện thuận lợi cho việc phát triển các chiến lược kinh doanh để nâng cao quản lý quan hệ khách hàng. Hơn nữa, hãy xem xét một công ty tư vấn có số lượng lớn các dự án. Để cải thiện quản lý dự án, có thể áp dụng phân cụm để phân chia các dự án thành các danh mục dựa trên mức độ tương đồng để có thể tiến hành kiểm toán và chẩn đoán dự án (để cải thiện việc phân phối và kết quả của dự án) 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 các hệ thống nhận dạng ký tự viết tay. Giả sử chúng ta có một tập dữ liệu các chữ số viết tay, trong đó mỗi chữ số được dán nhãn là 1, 2, 3, v.v. Lưu ý rằng có thể có sự khác biệt lớn về cách mọi người viết cùng một chữ số. Lấy số 2 làm ví dụ. Một số người có thể viết nó bằng một vòng tròn nhỏ ở phần dưới bên trái, trong khi một số khác thì không. Chúng ta có thể sử dụng phân cụm để xác định các lớp con cho "2", mỗi lớp biểu thị một biến thể về cách viết số 2. Sử dụng nhiều mô hình dựa trên các lớp con có thể cải thiện độ chính xác nhận dạng tổng thể.

Phân cụm cũng đã tìm thấy nhiều ứng dụng trong tìm kiếm trên Web. Ví dụ, tìm kiếm từ khóa thường có thể trả về số lượng kết quả rất lớn (tức là các trang có liên quan đến tìm kiếm) do số lượng trang web cực lớn. Phân cụm có thể được sử dụng để sắp xếp kết quả tìm kiếm thành các nhóm và trình bày kết quả theo cách ngắn gọn và dễ truy cập.

Hơn nữa, các kỹ thuật phân nhóm đã được phát triển để phân nhóm các tài liệu thành các chủ đề, thường được sử dụng 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 có được cái nhìn sâu sắc về phân phối dữ liệu, để quan sát các đặc điểm của từng cụm và tập trung vào một tập hợp cụm cụ thể để phân tích thêm. Ngoài ra, nó có thể đóng vai trò là bước tiền xử lý cho các thuật toán khác, chẳng hạn như đặc tính, lựa chọn tập hợp thuộc tính và phân loại, sau đó sẽ hoạt động trên các cụm đã phát hiện và các thuộc tính hoặc tính năng đã chọn.

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

Phân cụm cũng được gọi là phân đoạn dữ liệu trong một số ứng dụng vì phân cụm phân chia các tập dữ liệu lớn thành các nhóm theo mức độ tương đồng của chúng. Phân cụm cũng có thể được sử dụng để phát hiện giá trị ngoại lệ, trong đó các giá trị ngoại lệ (các giá trị "xa" bất kỳ cụm nào) có thể thú vị hơn các trường hợp thông thường. Các ứng dụng của phát hiện giá trị 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 các giao dịch mua rất đắt tiền và không thường xuyên, có thể được quan tâm như các hoạt động gian lận có thể xảy ra. Phát hiện giá trị 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 rất 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 phần mềm hoặc hệ thống phân tích thống kê, chẳng hạn như S-Plus, SPSS và SAS. Trong học máy, hãy nhớ rằng phân loại được gọi là học có giám sát vì thông tin nhãn lớp được cung cấp, nghĩa là thuật toán học được giám sát theo nghĩa là nó được thông báo về tư cách thành viên lớp của mỗi bộ dữ liệu đào tạo. 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 các cơ sở dữ liệu lớn. Các chủ đề nghiên cứu tích cực tập trung vào khả năng mở rộng của các phương pháp phân cụ, hiệu quả của các phương pháp phân cụ các hình dạng phức tạp (ví dụ: không lồi) và các loại dữ liệu (ví dụ: văn bản, đồ thị và hình ảnh), các kỹ thuật phân cụ đa chiều (ví dụ: phân cụ các đối tượng có hàng nghìn đặc điểm) và các phương pháp phân cụ dữ liệu số và danh nghĩa hỗn hợp trong cơ sở dữ liệu lớn.

10.1.2 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 việc 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 tình huống tìm kiếm trên Web. Phân cụm chỉ trên một mẫu của một tập dữ liệu lớn nhất định có thể dẫn đến kết quả sai lệch. Do đó, cần có các thuật toán phân cụm có khả năng mở rộng cao.
- **Khả năng xử lý các loại thuộc tính khác nhau:** Nhiều thuật toán được thiết kế để nhóm dữ liệu số (dựa trên khoảng). Tuy nhiên, các ứng dụng có thể yêu cầu nhóm các loại dữ liệu khác, chẳng hạn như dữ liệu nhị phân, danh nghĩa (phân loại) và thứ tự hoặc hỗn hợp các loại dữ liệu này. Gần đây, ngày càng nhiều ứng dụng cần các kỹ thuật nhóm cho các loại 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. Ví dụ, hãy xem xét các cảm biến thường được triển khai để giám sát môi trường. Phân tích cụm trên các phép đo 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 lan rộng, thường không có hình cầu. Điều quan trọng là phải 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 tham số đầu vào:** Nhiều thuật toán phân cụm yêu cầu người dùng cung cấp kiến thức miền dưới dạng 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. Các tham số thường khó xác định, đặc biệt là đối với các tập dữ liệu có nhiều chiều và khi người dùng vẫn chưa nắm bắt được hiểu biết sâu sắc về dữ liệu của họ. Việc yêu cầu chỉ định kiến thức miền không chỉ gây gánh nặng cho người dùng mà còn khiến chất lượng phân cụm khó kiểm soát.
- **Khả năng xử lý dữ liệu nhiễu:** Hầu hết các tập dữ liệu thực tế đều chứa các giá trị ngoại lai và/or dữ liệu bị thiếu, không xác định hoặc sai. Ví dụ, các phép đo cảm biến thường bị nhiễu-một số phép đo có thể không chính xác do cơ chế cảm biến và một số phép đo có thể 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 giá 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 bản cập nhật giá tăng vào các cấu trúc phân cụm hiện có và thay vào đó, phải tính toán lại một phân cụm mới từ đầu. Các thuật toán phân cụm cũng có thể nhạy cảm với thứ tự dữ liệu đầu vào. Nghĩa là, với một tập hợp các đối tượng dữ liệu, các thuật toán phân cụm có thể trả về các phân cụm khác nhau đáng kể tùy thuộc vào thứ tự mà các đối tượng được trình bày. Cần có các thuật toán phân cụm gia tăng và các thuật toán không nhạy cảm với thứ tự đầu vào.

- **Khả năng phân cụm dữ liệu có nhiều chiều:** Một tập dữ liệu có thể chứa nhiều chiều hoặc thuộc tính. Ví dụ, khi phân cụm tài liệu, mỗi từ khóa có thể được coi là một chiều và thường có hàng nghìn từ khóa. Hầu hết các thuật toán phân cụm đều xử lý tốt dữ liệu có chiều thấp như các tập dữ liệu chỉ bao gồm hai hoặc ba chiều. Việc tìm cụm các đối tượng dữ liệu trong không gian có nhiều chiều là một thách thức, đặc biệt là khi dữ liệu như vậy có thể rất thưa thớt và bị lệch rất nhiều.
- **Phân cụm dựa trên ràng buộc:** Các ứng dụng trong thế giới thực có thể cần thực hiện phân cụm theo nhiều loại ràng buộc khác nhau. Giả sử ràng công việc của bạn là chọn vị trí cho một số lượng máy ATM mới nhất định trong một thành phố. Để quyết định về điều này, bạn có thể phân cụm các hộ gia đình trong khi xem xét các ràng buộc như sông ngòi và mạng lưới 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 cụm có thể diễn giải, dễ hiểu và có thể sử dụng. Nghĩa là, phân cụm có thể cần được liên kết với các ứng dụng và diễn giải ngữ nghĩa a cụ thể. Điều quan trọng là phải nghiên cứu cách mục tiêu ứng dụng có thể ảnh hưởng đến việc lựa chọn các tính năng phân cụm và phương pháp phân cụm.

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

- **Tiêu chí phân vùng:** Trong một số phương pháp, tất cả các đối tượng được phân vùng sao cho không có hệ thống phân cấp nào tồn tại giữa các cụm. Nghĩa là, tất cả các cụm đều ở cùng một cấp độ khái niệm. Một phương pháp như vậy hữu ích, ví dụ, để phân vùng khách hàng thành các nhóm sao cho mỗi nhóm có người quản lý riêng. Ngoài ra, các phương pháp khác phân vùng các đối tượng dữ liệu theo thứ bậc, trong đó các cụm có thể được hình thành ở các cấp độ ngữ nghĩa khác nhau. Ví dụ, trong khai thác văn bản, chúng ta có thể muôn sắp xếp một tập hợp các tài liệu thành nhiều chủ đề chung, chẳng hạn như "chính trị" và "thể thao", mỗi chủ đề có thể có các chủ đề phụ. Ví dụ, "bóng đá", "bóng rổ", "bóng chày" và "khúc côn cầu" có thể tồn tại dưới dạng các chủ đề phụ của "thể thao". Bốn chủ đề phụ sau ở cấp độ thấp hơn trong hệ thống phân cấp so với "thể thao".
- **Phân tách cụm:** Một số phương pháp phân vùng các đối tượng dữ liệu thành các cụm loại trừ lẫn nhau. Khi phân cụm khách hàng thành các nhóm sao cho mỗi nhóm được một người quản lý chăm sóc, mỗi khách hàng chỉ có thể thuộc về một nhóm. Trong một số tình huống khác, các cụm có thể không loại trừ, nghĩa là một đối tượng dữ liệu có thể thuộc về nhiều cụm. Ví dụ, khi phân cụm 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 loại trừ.
- **Đo lường độ tương đồng:** Một số phương pháp xác định độ tương đồng giữa hai đối tượng theo khoảng cách giữa chúng. Khoảng cách như vậy có thể được xác định trên không gian Euclid,

mạng lưới đường bộ, không gian vectơ hoặc bất kỳ không gian nào khác. Trong các phương pháp khác, sự tương đồng có thể được xác định bằng kết nối dựa trên mật độ hoặc tính liên tục và có thể không dựa vào khoảng cách tuyệt đối giữa hai đối tượng. Các biện pháp tương đồng đóng vai trò cơ bản trong 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.

- Không gian cụm: Nhiều phương pháp cụm tìm kiếm các cụm trong toàn bộ không gian dữ liệu đã cho. Các phương pháp này hữu ích cho các tập dữ liệu có chiều thấp. Tuy nhiên, với dữ liệu có chiều cao, có thể có nhiều thuộc tính không liên quan, có thể khiến các phép đo độ tương đồng không đáng tin cậy. Do đó, các cụm được tìm thấy trong toàn bộ không gian thường không có ý nghĩa. Thay vào đó, thường tốt hơn là tìm kiếm các cụm trong các không gian con khác nhau của cùng một tập dữ liệu. Cụm không gian con khám phá các cụm và không gian con (thường có chiều thấp) thể hiện sự tương đồng của đối tượng.

Tóm lại, các thuật toán phân cụm có một số yêu cầu. Các yếu tố này bao gồm khả năng mở rộng và khả năng xử lý các loại thuộc tính khác nhau, dữ liệu nhiều, 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 về mức phân vùng, các cụm có loại trừ lẫn nhau hay không, các biện pháp tương tự được sử dụng và có thực hiện phân cụm không gian con hay không.

10.1.3 Tổng quan về các phương pháp phân cụm cơ bản

Có nhiều thuật toán phân cụm trong tài liệu.Ật khó để cung cấp một danh mục 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 nhiều danh mục. Tuy nhiên, việc trình bày một bức tranh tổng quát có thể chức về các phương pháp phân cụm là hữu ích. Nhìn chung, các phương pháp phân cụm cơ bản chính có thể được phân loại thành các danh mục sau, được thảo luận trong phần còn lại của chương này.

Phương pháp phân vùng: Với một tập hợp n đối tượng, phương pháp phân vùng sẽ xây dựng k phân vùng dữ liệu, trong đó mỗi phân vùng biểu diễn một cụm và $k \leq n$. Nghĩa là, nó chia dữ liệu thành k nhóm sao cho mỗi nhóm phải chứa ít nhất một đối tượng.

Nói cách khác, phương pháp phân vùng thực hiện phân vùng một cấp trên các tập dữ liệu. Các phương pháp phân vùng cơ bản thường áp dụng phân tách cụm độc quyền. Nghĩa là, mỗi đối tượng phải thuộc về chính xác một nhóm. Yêu cầu này có thể được nới lỏng, ví dụ, trong các kỹ thuật phân vùng mờ. Tài liệu tham khảo về các kỹ thuật như vậy được đưa ra trong các ghi chú thư mục (Phần 10.9).

Hầu hết các phương pháp phân vùng đều dựa trên khoảng cách. Với k , số lượng phân vùng cần xây dựng, phương pháp phân vùng sẽ tạo ra một phân vùng ban đầu. Sau đó, nó sử dụng kỹ thuật di dời lặp đi lặp lại nhằm cải thiện phân vùng bằng cách di chuyển các đối tượng từ nhóm này sang nhóm khác. Tiêu chí chung của một phân vùng tốt là các đối tượng trong cùng một cụm phải "gần" hoặc có liên quan với nhau, trong khi các đối tượng trong các cụm khác nhau phải "xã nhau" hoặc rất khác nhau. Có nhiều loại phân vùng khác

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

Việc đạt được tính tối ưu toàn cục trong phân cụm dựa trên phân vùng thường là điều cấm kỵ về mặt tính toán, có khả năng yêu cầu phải liệt kê đầy đủ tất cả các phân vùng có thể. 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 thuật toán k-means và k-medoids, giúp cải thiện dần chất lượng phân cụm và tiếp cận mức tối ưu cục bộ. Các phương pháp phân cụm heuristic này hoạt động tốt để tìm các cụm hình cầu trong các cơ sở dữ liệu có kích thước từ nhỏ đến trung bình. Để tìm các cụm có hình dạng phức tạp và đối với các tập dữ liệu rất lớn, cần phải mở rộng các phương pháp dựa trên phân vùng. Các phương pháp phân cụm dựa trên phân vùng được nghiên cứu sâu hơn trong Phần 10.2.

Phương pháp phân cấp: Phương pháp phân cấp tạo ra sự phân tích phân cấp của tập hợp các đối tượng dữ liệu đã cho. Phương pháp phân cấp có thể được phân loại là kết tụ hoặc chia tách, dựa trên cách phân tích phân cấp được hình thành.

Phương pháp tiếp cận kết tụ, còn được gọi là phương pháp 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 (mức cao nhất của hệ thống phân cấp), hoặc một điều kiện kết thúc được giữ nguyên. Phương pháp tiếp cận phân chia, còn được gọi là phương pháp tiếp cận 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 một 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. Nhiều phần mở rộng khác nhau của các phương pháp phân cấp cũng xem xét việc phân cụm trong các không gian con.

Các phương pháp phân cấp có nhược điểm là một khi một bước (hợp nhất hoặc tách) đã được thực hiện, nó sẽ không bao giờ có thể hoàn tác được. Sự cứng nhắc này hữu ích ở chỗ nó dẫn đến chi phí tính toán nhỏ hơn do không phải lo lắng về số lượng kết hợp của các lựa chọn khác nhau.

Các kỹ thuật như vậy không thể sửa các quyết định sai lầm; tuy nhiên, các phương pháp cải thiện chất lượng của cụm phân cấp đã được đề xuất. Các phương pháp cụm phân cấp được nghiên cứu trong Phần 10.3.

Các phương pháp dựa trên mật độ: Hầu hết các phương pháp phân vùng nhóm các đối tượng dựa trên khoảng cách giữa các đối tượng. Các phương pháp như vậy chỉ có thể tìm thấy các cụm hình cầu và gặp khó khăn trong việc khám phá các cụm có hình dạng tùy ý. Các phương pháp nhóm khác đã được phát triển dựa trên khái niệm mật độ. Ý tưởng chung của chúng là tiếp tục phát triển một cụm nhất định miễn là mật độ (số lượng đối tượng hoặc điểm dữ liệu) trong "vùng lân cận" vượt quá một ngưỡng nào đó. Ví dụ, đối với 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 nhiễu hoặc các giá trị ngoại lai và khám phá 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 cụm độc quyền hoặc một hệ thống phân cấp các cụm. Thông thường, các phương pháp dựa trên mật độ chỉ xem xét các cụm độc quyền và không xem xét các cụm mở. Hơn nữa, các phương pháp dựa trên mật độ có thể được mở rộng từ cụm không gian đầy đủ sang cụm không gian con. Các phương pháp cụm dựa trên mật độ được nghiên cứu trong Phần 10.4.

Các phương pháp dựa trên lưới: Các phương pháp dựa trên lưới lượng tử hóa không gian đối tượng thành một số lượng hữu hạn các ô tạo thành một cấu trúc lưới. Tất cả các hoạt động nhóm được thực hiện 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 cả phân cụm. Do đó, các phương pháp dựa trên lưới có thể được tích hợp với các phương pháp phân cụm khác như các phương pháp dựa trên mật độ và các 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 cụ, do đó đôi khi khó phân loại một thuật toán nhất định thành một thuật toán chỉ thuộc một loại phương pháp phân cụ duy nhất.

Hơn nữa, một số ứng dụng có thể có tiêu chí phân cụ đòi hỏi phải tích hợp nhiều kỹ thuật phân cụ.

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

Phương pháp	Đặc điểm chung
Phương pháp phân vùng	<ul style="list-style-type: none"> - Tìm các cụm hình cầu loại trừ lẫn nhau - Dựa trên khoảng cách - Có thể sử dụng mean hoặc medoid (v.v.) để biểu diễn trung tâm cụm - Hiệu quả đối với các tập dữ liệu có quy mô từ nhỏ đến trung bình
Phương pháp phân cấp	<ul style="list-style-type: none"> - Phân cụm là sự phân rã theo thứ bậc (tức là nhiều cấp độ) - Không thể sửa lỗi hợp nhất hoặc tách - Có thể kết hợp các kỹ thuật khác như phân cụm nhỏ hoặc xem xét các "liên kết" đối tượng
Phương pháp dựa trên mật độ	<ul style="list-style-type: none"> - Có thể tìm thấy các cụm có hình dạng tùy ý - Các cụm là các vùng vật thể dày đặc trong không gian được ngăn cách bởi các vùng có mật độ thấp - Mật độ cụm: Mỗi điểm phải có số lượng tối thiểu điểm trong "khu vực lân cận" của nó - Có thể lọc ra các giá trị ngoại lệ
Phương pháp dựa trên lưới	<ul style="list-style-type: none"> - Sử dụng cấu trúc dữ liệu lưới đa độ 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 đối tượng d chiều. Các đối tượng được biểu diễn bằng phông chữ in nghiêng đậm (ví dụ: p).

10.2 Phương pháp phân vùng

Phiên bản đơn giản nhất và cơ bản nhất của phân tích cụm là phân vùng, sắp xếp các đối tượng của một tập hợp thành nhiều 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 dưới dạng kiến thức nền. Tham số này là điểm khởi đầu cho các phương pháp phân vùng.

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

Trong phần này, bạn sẽ tìm hiểu các phương pháp phân vùng phổ biến và được biết đến nhiều nhất – k-means (Phần 10.2.1) và k-medoids (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.

10.2.1 k-Means: Một kỹ thuật dựa trên tâm

Giả sử một tập dữ liệu, D, chứa n đối tượng trong không gian Euclid. Các phương pháp phân vùng phân phối các đối tượng trong D thành k cụm, C_1, \dots, C_k , tức là, $C_i \subseteq D$ và $C_i \cap C_j = \emptyset$ với $(1 \leq i, j \leq k)$. Một hàm mục tiêu được sử dụng để đánh giá chất lượng phân vùng sao cho các đối tượng trong một cụm giống nhau nhưng không giống với các đối tượng trong các cụm khác. Nghĩa là, hàm mục tiêu hướng đến độ tương đồng nội cụm cao và độ tương đồng liên cụm thấp.

Kỹ thuật phân vùng dựa trên tâm sử dụng tâm của cụm, C_i , để biểu diễn cụm đó. Về mặt khái niệm, tâm của cụm là điểm trung tâm của cụm. Tâm có thể được định nghĩa theo nhiều cách khác nhau như theo giá trị trung bình hoặc medoid của các đối tượng (hoặc điểm) được gán cho cụm. Sự khác biệt giữa một đối tượng p và C_i đại diện của cụm, được đo bằng $dist(p, c_i)$, trong đó $dist(x, y)$ là khoảng cách Euclid giữa hai điểm x và y. Chất lượng của cụm C_i có thể được đo bằng biến thiên trong cụm, là tổng bình phương sai số giữa tất cả các đối tượng trong C_i và tâm c_i , được định nghĩa là

$$E = \sum_{i=1}^n dist(p, c_i)^2 \quad (10.1)$$

trong đó E là tổng của lỗi 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 c_i là tâm của cụm C_i (cả p và c_i đều là đa chiều). Nói cách khác, đối với mỗi đối tượng trong mỗi cụm, khoảng cách từ

đối tượng đến tâm cụm của nó được bình phong và các khoảng cách được cộng lại. Hàm mục tiêu này cố gắng làm cho k cụm kết quả trở nên nhỏ gọn và tách biệt nhất có thể.

Tối ưu hóa biến thiên trong cụm là một thách thức về mặt tính toán. Trong trường hợp xấu nhất, chúng ta sẽ phải liệt kê một số phân vùng có thể có theo cấp số nhân với số cụm và kiểm tra các giá trị biến thiên trong cụm. Người ta đã chỉ ra rằng bài toán này là NP-khó trong không gian Euclid nói chung ngay cả đối với hai cụm (tức là $k = 2$). Hơn nữa, bài toán này là NP-khó đối với số cụm nói chung k ngay cả trong không gian Euclid 2 chiều. Nếu số cụm k và số chiều của không gian $d \geq k+1$ là cố định, thì bài toán có thể được giải trong thời gian $O(n^d)$. Để khắc phục chi phí tính toán quá cao đối với giải pháp chính xác, các phương pháp tiếp cận tham lam thường được sử dụng trong thực tế. Một ví dụ điển hình là thuật toán k-means, thuật toán này đơn giản và thường được sử dụng.

"Thuật toán k-means hoạt động như thế nào?" Thuật toán k-means định nghĩa trung tâm của một cụm là giá trị trung bình của các điểm trong cụm. Thuật toán tiến hành như sau. Đầu tiên, thuật toán chọn ngẫu nhiên k đối tượng trong D, mỗi đối tượng ban đầu biểu diễn một trung bình hoặc tâm cụm. Đối với mỗi đối tượng còn lại, một đối tượng được gán cho cụm mà nó giống nhất, dựa trên khoảng cách Euclid giữa đối tượng và trung bình cụm. Sau đó, thuật toán k-means cải thiện theo từng bước biến thiên 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. Sau đó, tất cả các đối tượng được gán lại bằng cách sử dụng giá trị trung bình đã cập nhật làm trung tâm cụm mới. Các lần lặp tiếp tục cho đến khi việc gán ổn định, nghĩa là các cụm được hình thành trong vòng hiện tại giống với các cụm được hình thành trong vòng trước. Quy trình k-means được tóm tắt trong Hình 10.2.

Thuật toán: k-means. Thuật toán k-means để phân vùng, trong đó trung tâm của mỗi cụm được biểu thị bằng giá trị trung bình của các đối tượng trong cụm.

Đầu vào:

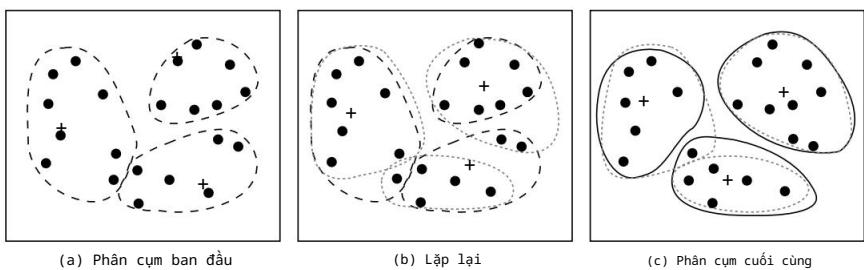
- k: số lượng cụm,
- D: một tập dữ liệu chứa n đối tượng.

Đầu ra: Một tập hợp k cụm.

Phương pháp:

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

Hình 10.2 Thuật toán phân vùng k-means.



Hình 10.3 Phân cụm một tập hợp các đối tượng sử dụng phương pháp k-means, đối với (b) cập nhật các trung tâm cụm và chỉ định 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 bằng phân vùng k-means. Xem xét một tập hợp các đối tượng nằm trong không gian 2 chiều, như được mô tả trong Hình 10.3(a). Giả sử $k = 3$, tức là người dùng muốn các đối tượng được phân cụm thành ba cụm.

Theo thuật toán trong Hình 10.2, chúng tôi tùy ý chọn ba đối tượng làm ba trung tâm cụm ban đầu, trong đó các trung tâm cụm được đánh dấu bằng dấu +. Mỗi đối tượng được gán cho một cụm dựa trên trung tâm cụm mà nó gần nhất. Phân phối như vậy tạo thành các hình bóng được bao quanh bởi các đường cong chấm bi, như thể hiện 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 gần nhất. Sự phân phối lại như vậy tạo thành những hình bóng mới được bao quanh bởi các đường cong đứt nét, như thể hiện trong Hình 10.3(b).

Quá trình này lặp lại, dẫn đến Hình 10.3(c). Quá trình gán lại các đối tượng theo từng cụm để cải thiện phân vùng được gọi là di dời theo từng cụm. Cuối cùng, không có sự gán lại nào của 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 pháp k-means không 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 lựa chọn ngẫu nhiên ban đầu của các trung tâm cụm. (Bạn sẽ được yêu cầu đưa ra một ví dụ để chứng minh điều này như một bài tập.) Để có được kết quả tốt trong thực tế, người ta thường chạy thuật toán k-means nhiều lần với các tâm cụm ban đầu khác nhau.

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

Có một số biến thể của phương pháp k-means. Chúng có thể khác nhau về cách lựa chọn k-means ban đầu, cách tính độ khác biệt và các chiến lược tính toán trung bình cụm.

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

Sự cần thiết để người dùng chỉ định k , số lượng cụm, trước có thể được coi là bất lợi. Đã có những nghiên cứu về cách khắc phục khó khăn này, tuy nhiên, chẳng hạn như như bằng cách cung cấp một phạm vi giá trị k gần đúng, và sau đó sử dụng một kỹ thuật phân tích để xác định k tốt nhất bằng cách so sánh các kết quả phân cụm thu được cho k khác nhau giá trị. Phương pháp k-means không phù hợp để khám phá các cụm có không lồi hình dạng hoặc cụm có kích thước rất khác nhau. Hơn nữa, nó nhạy cảm với tiếng ồn và dữ liệu ngoại lệ điểm vì một số lượng nhỏ dữ liệu như vậy có thể ảnh hưởng đáng kể đến giá trị trung bình.

"Làm thế nào chúng ta có thể làm cho thuật toán k-means có khả năng mở rộng hơn?" Một cách tiếp cận để làm cho phương pháp k-means hiệu quả hơn trên các tập dữ liệu lớn là sử dụng một tập hợp có kích thước tốt mẫu trong cụm. Một cách khác là sử dụng phương pháp lọc sử dụng chỉ số dữ liệu phân cấp không gian để tiết kiệm chi phí khi tính toán phương tiện. Một cách tiếp cận thứ ba khám phá ý tưởng về cụm vi mô, đầu tiên nhóm các vật thể gần nhau thành "cụm vi mô" và sau đó thực hiện phân cụm k-means trên các cụm vi mô. Phân cụm vi mô được thảo luận thêm trong Phần 10.3.

10.2.2 k-Medoids: Một kỹ thuật dựa trên đối tượng tiêu biểu

Thuật toán k-means nhạy cảm với các giá trị ngoại lai vì các đối tượng như vậy ở rất xa phần lớn dữ liệu và do đó, khi được gán vào một cụm, chúng có thể làm biến dạng đáng kể giá trị trung bình của cụm. Điều này vô tình ảnh hưởng đến việc chỉ định các đối tượng khác đến các cụm. Hiệu ứng này đặc biệt trầm trọng hơn do sử dụng lỗi bình phương hàm của Phương trình (10.1), như được quan sát trong Ví dụ 10.2.

Ví dụ 10.2 Một nhược điểm của k-means. Xem xét sáu điểm trong không gian 1 chiều có các giá trị

$1, 2, 3, 8, 9, 10$ và 25 tương ứng. Theo trực giác, bằng cách quan sát trực quan, chúng ta có thể tưởng tượng 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 ngoại lệ. k-means sẽ phân chia các giá trị như thế nào? Nếu chúng ta áp dụng k-means sử dụng $k = 2$ và Eq. (10.1), phân vùng $\{\{1, 2, 3\}, \{8, 9, 10, 25\}\}$ có sự thay đổi 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 giá trị trung bình của cụm $\{1, 2, 3\}$ là 2 và giá trị trung bình của $\{8, 9, 10, 25\}$ là 13. So sánh điều này với phân vùng $\{\{1, 2, 3, 8\}, \{9, 10, 25\}\}$, trong đó k-means tính toán biến thể trong cụm như

$$\begin{aligned} & (1 - 3,5)^2 + (2 - 3,5)^2 + (3 - 3,5)^2 + (8 - 3,5)^2 + (9 - 14,67)^2 \\ & + (10 - 14,67)^2 + (25 - 14,67)^2 = 189,67, \end{aligned}$$

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 sau có sự thay đổi trong cụm thấp nhất; do đó, phương pháp k-means 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, cách xa đáng kể so với tất cả các thành viên trong cụm.

"Làm thế nào chúng ta có thể sửa đổi thuật toán k-means để giảm độ nhạy cảm với các giá trị ngoại lai?" 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 biểu diễn 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 giống nhất. Sau đó, phương pháp phân vùng được thực hiện dựa trên nguyên tắc giảm thiểu tổng các điểm khác biệt giữa mỗi đối tượng p và đối tượng đại diện tương ứng của nó. Nghĩa là, tiêu chuẩn lỗi tuyệt đối được sử dụng, được định nghĩa là

$$E = \sum_{i=1}^n \text{dist}(p, o_i), \quad (10.2)$$

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

Khi $k = 1$, ta có thể tìm được trung vị chính xác trong $O(n^2)$ thời gian. Tuy nhiên, khi k là một dương tổng quát, bài toán k-medoid là NP-khó.

Thuật toán Phân vùng quanh Medoids (PAM) (xem Hình 10.5 sau) là một hiện thực hóa phổ biến của cụm k-medoids. Nó giải quyết vấn đề theo cách lặp đi lặp lại, tham lam. Giống như thuật toán k-means, các đối tượng đại diện ban đầu (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 đại diện có cải thiện chất lượng cụm hay không. Tất cả các thay thế có thể đều được thử. Quá trình lặp đi lặp lại của việc thay thế các đối tượng đại diện bằng các đối tượng khác tiếp tục cho đến khi chất lượng của cụm kết quả không thể được cải thiện bằng bất kỳ sự thay thế nào. Chất lượng này được đo bằng hàm chi phí của sự khác biệt trung bình giữa một đối tượng và đối tượng đại diện của cụm của nó.

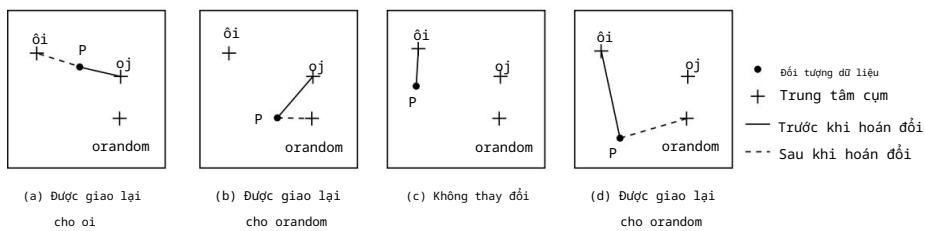
Cụ thể, hãy cho o_1, \dots, o_k là tập hợp các đối tượng đại diện hiện tại (tức là medoids).

Để xác định xem một đối tượng không đại diện, được biểu thị bằng o_{random} , có phải là sự thay thế tốt cho medoid o_j hiện tại ($1 \leq j \leq k$) hay không, chúng ta tính khoảng cách từ mọi đối tượng p đến đối tượng gần nhất trong tập $\{o_1, \dots, o_{j-1}, o_{random}, o_{j+1}, \dots, o_k\}$ và sử dụng khoảng cách để cập nhật hàm chi phí. Việc gán lại các đối tượng vào $\{o_1, \dots, o_{j-1}, o_{random}, o_{j+1}, \dots, o_k\}$ rất đơn giản. Giả sử đối tượng p hiện được gán cho một cụm được biểu diễn bằng medoid o_j (Hình 10.4a hoặc b). Chúng ta có cần gán lại p cho một cụm khác nếu o_j đang được thay thế bằng o_{random} không? Đối tượng p cần được gán lại thành o_{random} hoặc một cụm khác được biểu diễn bằng o_i ($i = j$), tùy theo cụm nào gần nhất.

Ví dụ, trong Hình 10.4(a), p gần nhất với o_1 và do đó được gán lại cho o_1 .

Tuy nhiên, trong Hình 10.4(b), p gần nhất với o_{random} và do đó được gán lại thành o_{random} . Nếu thay vào đó, p hiện được gán cho một cụm được biểu diễn bởi một số đối tượng khác o_i , $i = j$ thì sao?

456 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản



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

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

Mỗi lần tái chỉ định xảy ra, một sự khác biệt về lỗi tuyệt đối, E, được đóng góp vào hàm chi phí. Do đó, hàm chi phí tính toán sự khác biệt về giá trị lỗi 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 bởi tất cả các đối tượng không đại diện. Nếu tổng chi phí là số âm, thì oj được thay thế hoặc hoán đổi bằng orandom vì lỗi tuyệt đối thực tế E bị giảm. Nếu tổng chi phí là số dương, thì đối tượng đại diện hiện tại được coi là chấp nhận được và không có gì thay đổi trong quá trình lặp lại. oj, "Phương pháp nào mạnh mẽ hơn k-means hay k-medoids?" Phương pháp k-medoids mạnh mẽ hơn k-means khi có nhiều và giá trị ngoại lai vì medoid ít bị ảnh hưởng bởi giá trị ngoại lai hoặc các giá trị cực đoan khác hơn giá trị trung bình. Tuy nhiên, độ phức tạp của mỗi lần lặp trong thuật toán k-medoids là $O(k(n - k))$ và k, việc tính toán như vậy trở nên rất tốn kém và tốn kém hơn nhiều so với phương pháp k-means. Cả hai phương pháp đều yêu cầu người dùng chỉ định k, tức là số k^2 . Đối với các giá trị lớn của n cụm.

"Làm thế nào chúng ta có thể mở rộng phương pháp k-medoids?" Một thuật toán phân vùng k-medoids diễn hình như PAM (Hình 10.5) hoạt động hiệu quả đối với các tập dữ liệu nhỏ, nhưng không mở rộng tốt đối với các tập dữ liệu lớn. Để xử lý các tập dữ liệu lớn hơn, có thể sử dụng phương pháp lấy mẫu có tên là CLARA (Clustering LARge Applications). Thay vì xem xét toàn bộ tập dữ liệu, CLARA sử dụng một mẫu ngẫu nhiên của tập dữ liệu. Sau đó, thuật toán PAM 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 sẽ hoạt động tốt nếu nó được tạo ra sao cho mỗi đối tượng có xác suất được chọn vào mẫu bằng nhau. Các đối tượng đại diện (medoids) được chọn có khả năng sẽ tương tự như các đối tượng được chọn từ toàn bộ tập dữ liệu. CLARA xây dựng các cụm từ nhiều mẫu ngẫu nhiên và trả về cụm tốt nhất làm đầu ra. Độ phức tạp của việc tính toán medoid trên một mẫu ngẫu nhiên là $O(ks^2 + k(n - k))$, 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 kích thước mẫu. Lưu ý rằng PAM tìm kiếm k-medoid tốt nhất trong một tập dữ liệu nhất định, trong khi CLARA tìm kiếm k-medoid tốt nhất trong mẫu được chọn của tập dữ liệu. CLARA không thể tìm thấy một cụm tốt nếu bất kỳ medoid nào được lấy mẫu tốt nhất lại cách xa k-medoid tốt nhất. Nếu một đối tượng

Thuật toán: k-medoids. 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 k cụm.

Phương pháp:

(1) tùy ý chọn k đối tượng trong D làm đối tượng đại diện ban đầu hoặc hạt giống; (2) lặp lại (3) chỉ định mỗi đối tượng còn lại vào cụm có đối tượng đại diện gần nhất; (4) chọn ngẫu nhiên một đối tượng không đại diện, orandom; (5) tính tổng chi phí, S, của việc hoán đổi đối tượng đại diện, oj, với orandom; (6) nếu $S < 0$ thì hoán đổi oj với orandom để tạo thành tập hợp k đối tượng đại diện mới; (7) cho đến khi không có thay đổi;

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

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

"Làm thế nào chúng ta có thể cải thiện chất lượng và khả năng mở rộng của CLARA?" Hãy nhớ rằng khi tìm kiếm các medoid tốt hơn, PAM sẽ 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 trong một mẫu ngẫu nhiên của tập dữ liệu. Một thuật toán ngẫu nhiên có tên là CLARANS (Phân cụm các ứng dụng lớn dựa trên Tìm kiếm ngẫu nhiên) đưa ra sự đánh đổi 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 các medoid hiện tại. Việc thay thế x bằng y có thể cải thiện tiêu chuẩn lỗi tuyệt đối không? Nếu có, thì việc thay thế được thực hiện. CLARANS tiến hành tìm kiếm ngẫu nhiên như vậy 1 lần. Tập hợp các medoid hiện tại sau 1 bước được coi là tối ưu cục bộ. CLARANS lặp lại quá trình ngẫu nhiên này m lần và trả về tối ưu cục bộ tốt nhất làm kết quả cuối cùng.

10.3 Phương pháp phân cấp

Trong khi các phương pháp phân vùng đáp ứng yêu cầu phân cụm 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 như trong một hệ thống phân cấp. Một phương pháp phân cụm theo hệ thống phân cấp hoạt động bằng cách nhóm các đối tượng dữ liệu thành một hệ thống phân cấp hoặc "cây" cụm.

Việc biểu diễn các đối tượng dữ liệu dưới dạng phân cấp rất hữu ích cho việc tóm tắt và trực quan hóa dữ liệu. Ví dụ, với tư cách là người quản lý nguồn nhân lực tại AllElectronics,

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

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

Trong các ví dụ trước, mặc dù chúng tôi phân vùng dữ liệu theo thứ bậc, chúng tôi không cho rằng dữ liệu có cấu trúc thứ bậc (ví dụ: quản lý ở cùng cấp trong hệ thống phân cấp AllElectronics của chúng tôi với nhân viên). Việc chúng tôi sử dụng hệ thống phân cấp ở đây chỉ 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 ta có thể tin rằng dữ liệu mang một cấu trúc phân cấp cơ bản mà chúng ta muốn khám phá. Ví dụ, phân cụm phân cấp có thể khám phá ra một hệ thống phân cấp cho nhân viên AllElectronics được cấu trúc theo, chẳng hạn, lương. Trong nghiên cứu về sự tiến hóa, phân cụm phân cấp có thể nhóm các loài động vật theo các đặc điểm sinh học của chúng để khám phá ra các con đường tiến hóa, là một hệ thống phân cấp của các loài. Một ví dụ khác, nhóm các cấu hình của một trò chơi chiến lược (ví dụ, cờ vua hoặc cờ đam) theo cách phân cấp có thể giúp phát triển các chiến lược trò chơi có thể được sử dụng để đào tạo người chơi.

Trong phần này, bạn sẽ nghiên cứu các phương pháp phân cụm theo thứ bậc. Phần 10.3.1 bắt đầu bằng thảo luận về phân cụm theo thứ bậc kết tụ so với phân chia, sắp xếp các đối tượng thành một thứ bậc bằng cách sử dụng chiến lược từ dưới lên hoặc từ trên xuống. Các phương pháp kết tụ-kết tụ bắt đầu với các đối tượng riêng lẻ dưới dạng các cụm, được hợp nhất theo từng phần để tạo thành các cụm lớn hơn. Ngược lại, các phương pháp phân chia ban đầu cho phép tách cả các đối tượng đã cho tạo thành một cụm, sau đó chúng chia thành các cụm nhỏ hơn theo từng phần.

Các phương pháp phân cụm phân cấp có thể gặp khó khăn liên quan đến việc lựa chọn điểm hợp nhất hoặc tách. Quyết định như vậy rất quan trọng, vì khi một nhóm đối tượng được hợp nhất hoặc tách, quy trình ở bước tiếp theo sẽ hoạt động trên các cụm mới được tạo.

Nó sẽ không hoàn tác những gì đã thực hiện trước đó, cũng không thực hiện hoán đổi đối tượng giữa các cụm. Do đó, các quyết định hợp nhất hoặc tách, nếu không được lựa chọn tốt, có thể dẫn đến các cụm chất lượng thấp. Hơn nữa, các phương pháp không mở rộng tốt vì mỗi quyết định hợp nhất hoặc tách cần phải kiểm tra và đánh giá nhiều đối tượng hoặc cụm.

Một hướng triền vọng để cải thiện chất lượng cụm của các phương pháp phân cấp là tích hợp cụm phân cấp với các kỹ thuật cụm khác, tạo ra cụm nhiều pha (hoặc đa pha). Chúng tôi giới thiệu hai phương pháp như vậy, cụ thể 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 các cấu trúc cây, trong đó các nút lá hoặc các nút không phải lá cấp thấp có thể được xem là "cụm vi mô" tùy thuộc vào thang độ phân giải. Sau đó, nó áp dụng các

thuật toán phân cụm để thực hiện phân cụm ví mô trên các phân cụm ví mô. 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 toán, phương pháp xác suất và phương pháp Bayesian. Các phương pháp kết tụ, chia tách và đa pha là thuật toán, nghĩa là chúng coi các đối tượng dữ liệu là xác định và tính toán các cụm theo khoảng cách xác định giữa các đối tượng. Các phương pháp xác suất sử dụng các mô hình xác suất để nắm bắt các cụm và đo lường chất lượng của các cụm theo độ phù hợp của các mô hình. Chúng tôi thảo luận về phân cụm phân cấp xác suất trong Phần 10.3.5. Các phương pháp Bayesian tính toán phân phối của các cụm có thể. Nghĩa là, thay vì đưa ra một cụm xác định duy nhất trên một tập dữ liệu, chúng trả về một nhóm các cấu trúc cụm và xác suất của chúng, tùy thuộc vào dữ liệu đã cho. Các phương pháp Bayesian được coi là một chủ đề nâng cao và không được thảo luận trong cuốn sách này.

10.3.1 Phân cụm phân cấp kết tụ so với phân chia

Phương pháp phân cụm phân cấp có thể là kết tụ hoặc phân chia, tùy thuộc vào việc phân tách phân cấp được hình thành theo cách từ dưới lên (sáp nhập) hay từ trên xuống (chia tách). Chúng ta hãy xem xét kỹ hơn các chiến lược này.

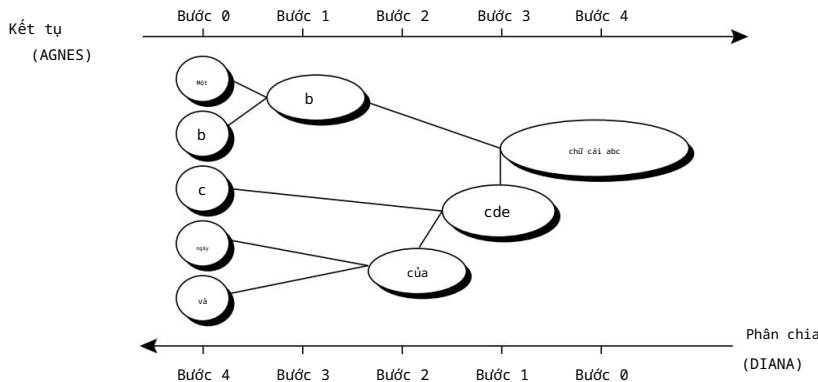
Phương pháp phân cụm phân cấp kết tụ sử dụng chiến lược từ dưới lên. Phương pháp này thường bắt đầu bằng cách để mỗi đối tượng hình thành cụm riêng của nó và hợp nhất các cụm theo từng bước thành các cụm ngày càng lớn hơn, cho đến khi tất cả các đối tượng nằm trong một cụm duy nhất hoặc các điều kiện kết thúc nhất định được đáp ứng. Cụm duy nhất trở thành gốc của hệ thống phân cấp. Đối với bước hợp nhất, phương pháp này tìm hai cụm gần nhau nhất (theo một số phép đo độ tương đồng) và hợp nhất hai cụm để tạo thành một cụm. Vì hai cụm được hợp nhất theo mỗi lần lặp, trong đó mỗi cụm chứa ít nhất một đối tượng, nên phương pháp kết tụ yêu cầu tối đa n lần lặp.

Phương pháp phân cụm 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 nhiều cụm con nhỏ hơn và phân vùng đệ quy các cụm đó thành các cụm nhỏ hơn. Quá trình phân vùng tiếp tục cho đến khi mỗi cụm ở cấp thấp nhất đủ mạch lạc hoặc chỉ chứa một đối tượng hoặc các đối tượng trong một cụm đủ giống nhau.

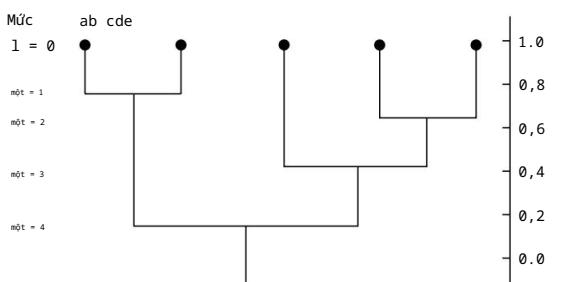
Trong cả phân cụm phân cấp kết tụ hoặc phân chia, người dùng có thể chỉ định số lượng cụm mong muốn làm điều kiện kết thúc.

Ví dụ 10.3 Phân cụm phân cấp theo kiểu kết tụ so với phân chia. Hình 10.6 cho thấy ứng dụng của AGNES (AGglomerative NESting), một phương pháp phân cụm phân cấp theo kiểu kết tụ, và DIANA (DIvisive ANALysis), một phương pháp phân cụm phân cấp theo kiểu chia, trên một tập dữ liệu gồm năm đối tượng, {a, b, c, d, e}. Ban đầu, AGNES, phương pháp kết tụ, đặt mỗi đối tượng vào một cụm của riêng nó. Sau đó, các cụm 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 trong C1 và một đối tượng trong C2 tạo thành khoảng cách Euclidean tối thiểu giữa bất kỳ hai đối tượng nào từ

460 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản



Hình 10.6 Phân cụm phân cấp kết tụ và phân chia trên các đôi tượng dữ liệu {a, b, c, d, e}.



Hình 10.7 Biểu diễn sơ đồ phân cấp cho cụm dữ liệu phân cấp của các đôi tượng dữ liệu {a, b, c, d, e}.

các cụm khác nhau. Đây là một phương pháp liên kết đơn trong đó mỗi cụm được biểu diễn bởi tất cả các đối tượng trong cụm và độ tương đồng giữa hai cụm được đo bằng độ tương đồng của cặp điểm dữ liệu gần nhất thuộc về các cụm khác nhau. Quá trình hợp nhất cụm lặp lại cho đến khi tất cả các đối tượng cuối cùng được hợp nhất để tạo thành một cụm.

DIANA, phương pháp phân chia, tiến hành theo cách tương phản. Tất cả các đối tượng được sử dụng để tạo thành một cụm ban đầu. Cụm được chia theo một số nguyên tắc như khoảng cách Euclidean tối đa giữa các đối tượng lân cận gần nhất trong cụm. Quá trình chia cụm lặp lại cho đến khi, cuối cùng, mỗi cụm mới chỉ chứa một đối tượng duy nhất.

Cấu trúc cây được gọi là sơ đồ cây phân nhánh thường được sử dụng để biểu diễn quá trình phân cụm phân cấp. Sơ đồ này cho thấy cách các đối tượng được nhóm lại với nhau (theo phương pháp kết tụ) hoặc phân vùng (theo phương pháp chia nhỏ) từng bước. Hình 10.7 cho thấy sơ đồ cây phân nhánh cho năm đối tượng được trình bày trong Hình 10.6, trong đó $l = 0$ cho thấy năm đối tượng là các cụm đơn lẻ ở mức 0. Tại $l = 1$, các đối tượng a và b được nhóm lại với nhau để tạo thành

cụm đầu tiên, và chúng ở cùng nhau ở tất cả các cấp độ tiếp theo. Chúng ta cũng có thể sử dụng trực đọc để hiển thị thang độ tương đồng giữa các cụm. Ví dụ, khi độ tương đồng của hai nhóm đối tượng, {a, b} và {c, d, e}, xấp xỉ bằng 0,16, chúng được hợp nhất với nhau để tạo thành một cụm duy nhất.

Một thách thức với các phương pháp chia tách là làm thế nào để phân chia một cụm lớn thành nhiều cụm nhỏ hơn. Ví dụ, có $2^n - 1$ cách có thể để phân chia một tập hợp n đối tượng thành hai tập hợp con loại trừ, trong đó n là số đối tượng. Khi n lớn, về mặt tính toán, việc kiểm tra tất cả các khả năng là điều cấm kỵ. Do đó, một phương pháp chia tách thường sử dụng phương pháp tìm kiếm trong phân vùng, điều này có thể dẫn đến kết quả không chính xác. Vì mục đích hiệu quả, các phương pháp chia tách thường không quay lại các quyết định phân vùng đã được đưa ra. Sau khi một cụm được phân vùng, bất kỳ phân vùng thay thế nào của cụm này sẽ không được xem xét lại. Do những thách thức trong các phương pháp chia tách, có nhiều phương pháp kết tụ hơn so với các phương pháp chia tách.

10.3.2 Đo khoảng cách trong các phương pháp thuật toán

Cho dù sử dụng phương pháp kết tụ hay phương pháp phân chia, nhu cầu cốt lõi là đo khoảng cách giữa hai cụm, trong đó mỗi cụm thường là một tập hợp các đối tượng.

Bốn biện pháp được sử dụng rộng rãi cho khoảng cách giữa các cụm như sau, trong đó $|p_i - p_j|$ là khoảng cách giữa hai đối tượng hoặc điểm, p_i và p_j ; m_i là giá trị trung bình của cụm, C_i ; và n_i là số đối tượng trong C_i . Chúng cũng được gọi là các biện pháp liên kết.

$$\text{Khoảng cách tối thiểu: } \text{distmin}(C_i, C_j) = \min_{C_j} \{ |p_i - p_j| \} \quad (10.3)$$

$$\text{Khoảng cách tối đa: } \text{distmax}(C_i, C_j) = \max_{C_j} \{ |p_i - p_j| \} \quad (10.4)$$

$$\text{Khoảng cách trung bình: } \text{distmean}(C_i, C_j) = |m_i - m_j| \quad (10.5)$$

$$\text{Khoảng cách trung bình: } \text{khoảng cách}(C_i, C_j) = \frac{1}{\min_{p_i \in C_i, p_j \in C_j}} \sum_{p_i, p_j} |p_i - p_j| \quad (10.6)$$

Khi một thuật toán sử dụng khoảng cách tối thiểu, $\text{dmin}(C_i, C_j)$, để đo khoảng cách giữa các cụm, đôi khi nó được gọi là thuật toán phân cụm láng giềng gần nhất. Hơn nữa, nếu quá trình phân cụm bị chấm dứt khi khoảng cách giữa các cụm gần nhất vượt quá ngưỡng do người dùng xác định, nó được gọi là thuật toán liên kết đơn. Nếu chúng ta xem các điểm dữ liệu là các nút của một đồ thị, với các cạnh tạo thành một đường dẫn giữa các nút trong một cụm, thì việc hợp nhất hai cụm, C_i và C_j , tương ứng với việc thêm một cạnh giữa cặp nút gần nhất trong C_i và C_j . Vì các cạnh liên kết các cụm luôn nằm giữa các cụm khác nhau, nên đồ thị kết quả sẽ tạo ra một cây. Do đó, một thuật toán phân cụm phân cấp kết tụ sử dụng phép đo khoảng cách tối thiểu cũng được gọi là

Thuật toán cây bao trùm tối thiểu, trong đó cây bao trùm của đồ thị là cây kết nối tất cả các đỉnh và cây bao trùm tối thiểu là cây có tổng trọng số cạnh nhỏ nhất.

Khi một thuật toán sử dụng khoảng cách tối đa, $d_{\max}(C_i, C_j)$, để đo khoảng cách giữa các cụm, đôi khi nó được gọi là thuật toán cụm hàng xóm xa nhất. Nếu quá trình cụm kết thúc khi khoảng cách tối đa giữa các cụm gần nhất vượt quá ngưỡng do người dùng xác định, nó được gọi là thuật toán liên kết hoàn chỉnh. Bằng cách xem các điểm dữ liệu như các nút của một đồ thị, với các cạnh liên kết các nút, chúng ta có thể coi mỗi cụm như một đồ thị con hoàn chỉnh, tức là, với các cạnh kết nối tất cả các nút trong các cụm. Khoảng cách giữa hai cụm được xác định bởi các nút xa nhất trong hai cụm.

Thuật toán lân cận xa nhất có xu hướng giảm thiểu sự gia tăng đường kính của các cụm tại mỗi lần lặp. Nếu các cụm thực sự khá nhỏ gọn và có kích thước gần bằng nhau, phương pháp này sẽ tạo ra các cụm chất lượng cao. Nếu không, các cụm được tạo ra có thể vô nghĩa.

Các phép đo tối thiểu và tối đa trước đây biểu thị hai thái cực trong việc đo khoảng cách giữa các cụm. Chúng có xu hướng quá nhạy cảm với các giá trị ngoại lai hoặc dữ liệu nhiễu. Việc sử dụng khoảng cách trung bình hoặc khoảng cách trung bình là sự thỏa hiệp giữa khoảng cách tối thiểu và tối đa và khắc phục được vấn đề nhạy cảm với giá trị ngoại lai. Trong khi khoảng cách trung bình là khoảng cách dễ tính toán nhất, thì khoảng cách trung bình có lợi thế ở chỗ nó có thể xử lý dữ liệu phân loại cũng như dữ liệu số. Việc tính toán vectơ trung bình cho dữ liệu phân loại có thể khó hoặc không thể xác định được.

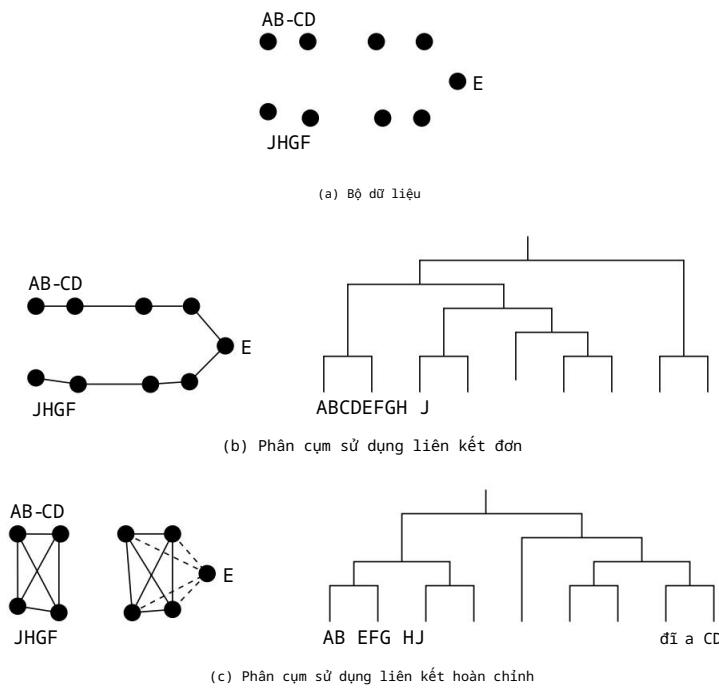
Ví dụ 10.4 Liên kết đơn so với liên kết hoàn chỉnh. Chúng ta hãy áp dụng phân cụm phân cấp cho tập dữ liệu của Hình 10.8(a). Hình 10.8(b) cho thấy sơ đồ phân cấp sử dụng liên kết đơn. Hình 10.8(c) cho thấy trường hợp sử dụng liên kết hoàn chỉnh, trong đó các cạnh giữa các cụm {A,B,J,H} và {C,D,G,F,E} được bỏ qua để dễ trình bày. Ví dụ này cho thấy rằng bằng cách sử dụng liên kết đơn, chúng ta có thể tìm thấy các cụm phân cấp được xác định bởi sự gần gũi cục bộ, trong khi liên kết hoàn chỉnh có xu hướng tìm thấy các cụm lựa chọn sự gần gũi toàn cục. ■

Có nhiều biến thể của bốn biện pháp liên kết thiết yếu vừa được thảo luận. Ví dụ, chúng ta có thể đo khoảng cách giữa hai cụm bằng khoảng cách giữa các tâm (tức là các đối tượng trung tâm) của cụm.

10.3.3 BIRCH: Phân cụm phân cấp đa pha sử dụng cây đặc điểm phân cụm

Balanced Iterative Reduction and Clustering using Hierarchies (BIRCH) được thiết kế để nhóm một lượng lớn dữ liệu số bằng cách tích hợp nhóm phân cấp (ở giai đoạn nhóm vi mô ban đầu) và các phương pháp nhóm khác như phân vùng lặp (ở giai đoạn nhóm vĩ mô sau này). Nó khắc phục được hai khó khăn trong các phương pháp nhóm tích tụ: (1) khả năng mở rộng và (2) không thể hoàn tác những gì đã thực hiện ở bước trước.

BIRCH sử dụng các khái niệm về tính năng cụm để tóm tắt một cụm và cây tính năng cụm (CF-tree) để biểu diễn một hệ thống phân cấp cụm. Các cấu trúc này giúp



Hình 10.8 Phân cụm theo thứ bậc sử dụng liên kết đơn và liên kết hoàn chỉnh.

Phương pháp phân cụ đạt được tốc độ và khả năng mở rộng tốt trong các cơ sở dữ liệu lớn hoặc thậm chí là cơ sở dữ liệu phát trực tuyến, đồng thời cũng hiệu quả trong việc phân cụ gia tăng và động các đối tượng đầu vào.

Hãy xem xét một cụm các đối tượng dữ liệu hoặc điểm n chiều. Tính năng cụm (CF) của cụm là một vector 3 chiều tóm tắt thông tin về các cụm đối tượng. Nó được định nghĩa là

$$CF = n, LS, SS, \quad (10.7)$$

trong đó LS là tổng tuyến tính của n điểm (tức là các $\sum_{i=1}^N xi$), và SS là tổng bình phương của điểm dữ liệu (tức là $\sum_{i=1}^N x_i^2$).

x_i . Tính năng cụm về cơ bản là bản tóm tắt các số liệu thống kê cho cụm đã cho.

Sử dụng tính năng phân cụm, chúng ta có thể dễ dàng suy ra nhiều số liệu thống kê hữu ích của một cụm. Ví dụ, tâm cụm, $x\theta$, bán kính, R và đường kính, D, là

$$x\theta = \frac{\sum_{i=1}^N xi}{N} = \frac{LS}{N}, \quad (10.8)$$

$$R = \frac{\overbrace{(x_i - x_0)^2}^{\text{tối}=1} \dots \overbrace{(x_N - x_0)^2}^{\text{tối}=N}}{N} = \frac{nSS}{\text{giá}^2} = \frac{2LS2 + nLS}{2LS2 + nLS}, \quad (10.9)$$

$$D = \frac{\overbrace{(x_i - x_j)^2}^{i=1 \dots j=1} \dots \overbrace{(x_n - x_1)^2}^{n(n-1)}}{n(n-1)} = \frac{2nSS}{n(n-1)} = \frac{2LS2}{2LS2 + nLS}. \quad (10.10)$$

Tại đây, R là khoảng cách trung bình từ các đối tượng thành viên đến tâm và D là khoảng cách trung bình theo cặp trong một cụm. Cả R và D đều phản ánh độ chật chẽ của cụm xung quanh tâm.

Tóm tắt một cụm sử dụng tính năng phân cụm có thể tránh lưu trữ thông tin chi tiết về các đối tượng hoặc điểm riêng lẻ. Thay vào đó, chúng ta chỉ cần một kích thước không gian không đổi để lưu trữ tính năng phân cụm. Đây là chìa khóa cho hiệu quả của BIRCH trong không gian. Hơn nữa, các tính năng cụm là cộng. Nghĩa là, đối với hai cụm rời rạc, $C1$ và $C2$, với các tính năng cụm $CF1 = n1, LS1, SS1$ và $CF2 = n2, LS2, SS2$, thì tính năng cụm cho cụm được hình thành bằng cách hợp nhất $C1$ và $C2$ chỉ đơn giản là

$$CF1 + CF2 = n1 + n2, LS1 + LS2, SS1 + SS2. \quad (10.11)$$

Ví dụ 10.5 Tính năng cụm. Giả sử có ba điểm, $(2,5), (3,2)$, và $(4,3)$, trong một cụm, $C1$. Tính năng cụm của $C1$ là

$$CF1 = 3, (2 + 3 + 4, 5 + 2 + 3), (2^{2 \cdot 2}, 52 + 2 + 3 + 4^{2 \cdot 2 + 3}) = 3, (9, 10), (29, 38).$$

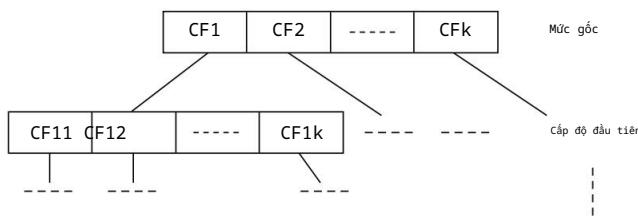
Giả sử $C1$ không liên quan đến cụm thứ hai, $C2$, trong đó $CF2 = 3, (35, 36), (417, 440)$.

Tính năng nhóm của một nhóm mới, $C3$, được hình thành bằng cách hợp nhất $C1$ và $C2$, được bắt nguồn bằng cách thêm $CF1$ và $CF2$. Nghĩa là,

$$CF3 = 3 + 3, (9 + 35, 10 + 36), (29 + 417, 38 + 440) = 6, (44, 46), (446, 478). \quad \blacksquare$$

Cây CF là cây cân bằng độ cao lưu trữ các tính năng cụm cho cụm phân cấp. Một ví dụ được hiển thị trong Hình 10.9. Theo định nghĩa, một nút không phải lá trong một cây có các nút con hoặc "con". Các nút không phải lá lưu trữ tổng CF của các nút con của chúng và do đó tóm tắt thông tin cụm về các nút con của chúng. Cây CF có hai tham số: hệ số phân nhánh, B và ngưỡng, T . Hệ số phân nhánh chỉ định số lượng nút con tối đa trên mỗi nút không phải lá. Tham số ngưỡng chỉ định đường kính tối đa của các cụm con được lưu trữ tại các nút lá của cây. Hai tham số này kiểm soát ngầm kích thước của cây kết quả.

Với một lượng bộ nhớ chính hạn chế, một cân nhắc quan trọng trong BIRCH là giảm thiểu thời gian cần thiết cho đầu vào/dầu ra (I/O). BIRCH áp dụng kỹ thuật cụm đa pha : Một lần quét duy nhất của tập dữ liệu tạo ra một cụm cơ bản, tốt và



Hình 10.9 Cấu trúc cây CF.

một hoặc nhiều lần quét bổ sung có thể được sử dụng tùy chọn để cải thiện chất lượng hơn nữa. Các giai đoạn chính là

- Giai đoạn 1: BIRCH quét cơ sở dữ liệu để xây dựng cây CF ban đầu trong bộ nhớ, có thể được xem như một quá trình nén dữ liệu đa cấp nhằm cố gắng bảo toàn cấu trúc cụm vốn có của dữ liệu.
- Giai đoạn 2: BIRCH áp dụng thuật toán phân cụ (được chọn) để phân cụ các nút lá của cây CF, giúp loại bỏ các cụm thừa thớt như các điểm ngoại lệ và nhóm các cụm dày đặc thành các cụm lớn hơn.

Đối với Giai đoạn 1, cây CF được xây dựng động khi các đối tượng được chèn vào. Do đó, phương pháp này là gia tăng. Một đối tượng được chèn vào mục lá gần nhất (cụm con). Nếu đường kính của cụm con được lưu trữ trong nút lá sau khi chèn lớn hơn giá trị ngưỡng, thì nút lá và có thể là các nút khác sẽ bị tách ra. Sau khi chèn đối tượng mới, thông tin về đối tượng được truyền đến gốc của cây. Kích thước của cây CF có thể được thay đổi bằng cách sửa đổi ngưỡng. Nếu kích thước bộ nhớ cần thiết để lưu trữ cây CF lớn hơn kích thước của bộ nhớ chính, thì có thể chỉ định giá trị ngưỡng lớn hơn và cây CF được xây dựng lại.

Quá trình xây dựng lại được thực hiện bằng cách xây dựng một cây mới từ các nút lá của cây cũ. Do đó, quá trình xây dựng lại cây được thực hiện mà không cần phải đọc lại tất cả các đối tượng hoặc điểm. Điều này tương tự như việc chèn và chia nút trong quá trình xây dựng cây B+. Do đó, để xây dựng cây, dữ liệu chỉ cần được đọc một lần. Một số phương pháp và thuật toán tìm kiếm đã được giới thiệu để xử lý các giá trị ngoại lai và cải thiện chất lượng của cây CF bằng cách quét dữ liệu bổ sung. Sau khi cây CF được xây dựng, bất kỳ thuật toán phân cụm nào, chẳng hạn như thuật toán phân vùng thông thường, đều có thể được sử dụng với cây CF trong Giai đoạn 2.

"BIRCH hiệu quả như thế nào?" Độ phức tạp thời gian của thuật toán là $O(n)$, trong đó n là số đối tượng cần nhóm lại. Các thí nghiệm đã chỉ ra khả năng mở rộng tuyến tính của thuật toán liên quan đến số lượng đối tượng và chất lượng nhóm dữ liệu tốt. Tuy nhiên, vì mỗi nút trong cây CF chỉ có thể chứa một số lượng mục nhập hạn chế do kích thước của nó, nên một nút cây CF không phải lúc nào cũng tương ứng với những gì người dùng có thể coi là một cụm tự nhiên. Hơn nữa, nếu các cụm không có hình cầu, BIRCH không hoạt động tốt vì nó sử dụng khái niệm bán kính hoặc đường kính để kiểm soát ranh giới của cụm.

Các ý tưởng về các tính năng cụm và cây CF đã được áp dụng ngoài BIRCH. Nhiều ý tưởng khác đã mượn để giải quyết các vấn đề về cụm luồng dữ liệu và dữ liệu động.

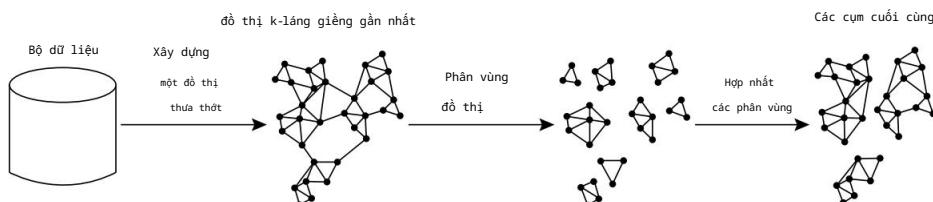
10.3.4 Chameleon: Phân cụm phân cấp đa pha Sử dụng mô hình động

Chameleon là một thuật toán phân cụm phân cấp sử dụng mô hình động để xác định độ tương đồng giữa các cặp cụm. Trong Chameleon, độ tương đồng của cụm được đánh giá dựa trên (1) mức độ kết nối tốt của các đối tượng trong một cụm và (2) mức độ gần nhau của các cụm. Nghĩa là, hai cụm được hợp nhất nếu khả năng kết nối của chúng cao và chúng ở gần nhau. Do đó, Chameleon không phụ thuộc vào mô hình tĩnh do người dùng cung cấp và có thể tự động thích ứng với các đặc điểm bên trong của các cụm đang được hợp nhất. Quá trình hợp nhất tạo điều kiện thuận lợi cho việc khám phá các cụm tự nhiên và đồng nhất và áp dụng cho tất cả các kiểu dữ liệu miễn là có thể chỉ định hàm tương đồng.

Hình 10.10 minh họa cách Chameleon hoạt động. Chameleon sử dụng phương pháp đồ thị k-nearest-neighbor để xây dựng đồ thị thừa thớt, trong đó mỗi đỉnh của đồ thị biểu diễn một đối tượng dữ liệu và tồn tại một cạnh giữa hai đỉnh (đối tượng) nếu một đối tượng nằm trong số k đối tượng giống nhất với đối tượng kia. Các cạnh được cân nhắc để phản ánh sự giống nhau giữa các đối tượng. Chameleon sử dụng thuật toán phân vùng đồ thị để phân vùng đồ thị k-nearest-neighbor thành một số lượng lớn các cụm con tương đối nhỏ sao cho nó giảm thiểu việc cắt cạnh. Nghĩa là, một cụm C được phân vùng thành các cụm con C_i và C_j để giảm thiểu trọng số của các cạnh sẽ bị cắt nếu C được chia đôi thành C_i và C_j . Nó đánh giá khả năng kết nối tuyệt đối giữa các cụm C_i và C_j .

Sau đó, Chameleon sử dụng thuật toán phân cụm phân cấp kết tụ lặp đi lặp lại hợp nhất các cụm con dựa trên mức độ tương đồng của chúng. Để xác định các cặp cụm con tương tự nhau, thuật toán này tính đến cả khả năng kết nối và mức độ gần nhau của các cụm. Cụ thể, Chameleon xác định mức độ giống nhau giữa mỗi cặp cụm C_i và C_j theo khả năng kết nối tương đối của chúng, $RI(C_i, C_j)$, và mức độ gần nhau tương đối của chúng, $RC(C_i, C_j)$.

- Sự kết nối tương đối, $RI(C_i, C_j)$, giữa hai cụm, C_i và C_j , được định nghĩa là sự kết nối tuyệt đối giữa C_i và C_j , được chuẩn hóa theo



Hình 10.10 Tắc kè hoa: phân cụm phân cấp dựa trên k-hàng xóm gần nhất và mô hình động.

Nguồn: Dựa trên Karypis, Han và Kumar [KHK99].

sự kết nối nội bộ của hai cụm, C_i và C_j . Nghĩa là, $|EC\{C_i, C_j\}| / (|ECC_i| + |ECC_j|)$

$$RI(C_i, C_j) = \frac{|ECC_j|}{\frac{|ECC_i| + |ECC_j|}{2}}, \quad (10.12)$$

trong đó $EC\{C_i, C_j\}$ là cạnh cắt như đã định nghĩa trước đó cho một cụm chứa cả C_i và C_j . Tương tự, ECC_i (hoặc ECC_j) là tổng nhỏ nhất của các cạnh cắt phân chia C_i (hoặc C_j) thành hai phần gần bằng nhau.

- Độ gần tương đối, $RC(C_i, C_j)$, giữa một cặp cụm, C_i và C_j , độ gần giữa C_i và C_j , là sự tuyệt đối được chuẩn hóa theo độ gần bên trong của hai cụm, C_i và C_j . Nó được định nghĩa là

$$RC(C_i, C_j) = \frac{\overline{SEC\{C_i, C_j\}}}{\frac{|C_i| - |C_j|}{|C_i| + |C_j| + SEC_{C_i} + SEC_{C_j}}}, \quad (10.13)$$

trong đó $\overline{SEC\{C_i, C_j\}}$ là trọng số trung bình của các cạnh kết nối các đỉnh trong C_i với các đỉnh trong C_j và SEC_{C_i} (hoặc SEC_{C_j}) là trọng số trung bình của các cạnh thuộc về phân giác cắt nhỏ của cụm C_i (hoặc C_j).

Chameleon đã được chứng minh là có sức mạnh lớn hơn trong việc phát hiện các cụm có hình dạng tùy ý có chất lượng cao so với một số thuật toán nổi tiếng như BIRCH và DBSCAN dựa trên mật độ (Phần 10.4.1). Tuy nhiên, chi phí xử lý cho các cụm có chiều cao 2^2 dữ liệu có thể yêu cầu $O(n^2)$ thời gian cho n đối tượng trong trường hợp xấu nhất.

10.3.5 Phân cụm phân cấp xác suất Các phương pháp phân cụm phân cấp

thuật toán sử dụng các phép đo liên kết có xu hướng dễ hiểu và thường hiệu quả trong phân cụm. Chúng thường được sử dụng trong nhiều ứng dụng phân tích phân cụm. Tuy nhiên, các phương pháp phân cụm phân cấp thuật toán có thể gặp phải một số nhược điểm. Đầu tiên, việc lựa chọn một phép đo khoảng cách tốt cho phân cụm phân cấp thường không hề đơn giản. Thứ hai, để áp dụng một phương pháp thuật toán, các đối tượng dữ liệu không thể có bất kỳ giá trị thuộc tính nào bị thiếu. Trong trường hợp dữ liệu được quan sát một phần (tức là một số giá trị thuộc tính của một số đối tượng bị thiếu), không dễ để áp dụng phương pháp phân cụm phân cấp thuật toán vì không thể thực hiện tính toán khoảng cách. Thứ ba, hầu hết các phương pháp phân cụm phân cấp thuật toán đều mang tính kinh nghiệm và tại mỗi bước, tìm kiếm cục bộ để có quyết định hợp nhất/tách tốt. Do đó, mục tiêu tối ưu hóa của phân cấp cụm kết quả có thể không rõ ràng.

Phân cụm phân cấp xác suất nhằm mục đích khắc phục một số nhược điểm này bằng cách sử dụng các mô hình xác suất để đo khoảng cách giữa các cụm.

Một cách để xem xét vấn đề phân cụm là coi tập hợp các đối tượng dữ liệu cần phân cụm là một mẫu của cơ chế tạo dữ liệu cơ bản cần phân tích hoặc, chính thức là mô hình tạo dữ liệu. Ví dụ, khi chúng tôi tiến hành phân tích phân cụm trên một tập hợp các cuộc khảo sát tiếp thị, chúng tôi giả định rằng các cuộc khảo sát được thu thập là một mẫu ý kiến của tất cả các khách hàng có thể. Ở đây, cơ chế tạo dữ liệu là một xác suất

phân phối ý kiến liên quan đến các khách hàng khác nhau, không thể thu thập trực tiếp và đầy đủ. Nhiệm vụ của việc phân cụm là ước tính mô hình tạo ra chính xác nhất có thể bằng cách sử dụng các đối tượng dữ liệu quan sát được để phân cụm.

Trong thực tế, chúng ta có thể cho rằng các mô hình tạo dữ liệu áp dụng các hàm phân phối chung, chẳng hạn như phân phối Gaussian hoặc phân phối Bernoulli, được điều chỉnh bởi các tham số. Nhiệm vụ học một mô hình tạo dữ liệu sau đó được giảm xuống thành việc tìm các giá trị tham số mà mô hình phù hợp nhất với tập dữ liệu quan sát được.

Ví dụ 10.6 Mô hình sinh. Giả sử chúng ta được cung cấp một tập hợp các điểm 1-D $X = \{x_1, \dots, x_n\}$ để phân tích cụm. Giả sử các điểm dữ liệu được tạo ra bởi phân phối Gaussian,

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (10.14)$$

trong đó các tham số là μ (trung bình) và σ^2 (phương sai).

Xác suất mà một điểm $x_i \in X$ sau đó được tạo ra bởi mô hình là

$$P(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (10.15)$$

Do đó, khả năng X được tạo ra bởi mô hình là

$$L(N(\mu, \sigma^2) : X) = P(X | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (10.16)$$

Nhiệm vụ của việc học mô hình sinh là tìm các tham số μ và σ^2 sao cho khả năng $L(N(\mu, \sigma^2) : X)$ được tối đa hóa, nghĩa là tìm thấy

$$N(\mu_0, \sigma_0^2) = \operatorname{argmax}\{L(N(\mu, \sigma^2) : X)\}, \quad (10.17)$$

trong đó $\max\{L(N(\mu, \sigma^2) : X)\}$ được gọi là khả năng lớn nhất. ■

Với một tập hợp các đối tượng, chất lượng của một cụm được hình thành bởi tất cả các đối tượng có thể được đo bằng khả năng tối đa. Đối với một tập hợp các đối tượng được phân chia thành m cụm C_1, \dots, C_m , chất lượng có thể được đo bằng

$$Q(\{C_1, \dots, C_m\}) = \prod_{i=1}^m P(C_i), \quad (10.18)$$

trong đó $P()$ là khả năng tối đa. Nếu chúng ta hợp nhất hai cụm, $Cj1$ và $Cj2$, cụm, vào một $Cj1 \cup Cj2$, thì, sự thay đổi về chất lượng của cụm tổng thể là

$$\begin{aligned} & Q(\{C_1, \dots, C_m\} \cup \{C_{j1}, C_{j2}\}) - Q(C_{j1} \cup C_{j2}) \\ &= \frac{\sum_{i=1}^{t_{j1}} P(C_i) \cdot P(C_{j1} \cup C_{j2})}{P(C_{j1})P(C_{j2})} - \frac{\sum_{i=1}^{t_{j1}} P(C_i)}{P(C_{j1})P(C_{j2})} \\ &= P(C_{j1}) \cdot \frac{P(C_{j1} \cup C_{j2})}{P(C_{j1})P(C_{j2})} - 1 \quad . \end{aligned} \quad (10.19)$$

Khi chọn hợp nhất hai cụm trong phân cụm phân cấp, $P(C_i)$ là hằng số ^m đối với bất kỳ cặp cụm nào. Do đó, với các cụm $C1$ và $C2$ đã cho, khoảng cách giữa chúng có thể được đo bằng

$$dist(C_i, C_j) = \log \frac{P(C_1 \cup C_2)}{P(C_1)P(C_2)} \quad (10.20)$$

Phương pháp phân cụ phân cấp xác suất có thể áp dụng khuôn khổ phân cụ kết tụ, nhưng sử dụng mô hình xác suất (Phương trình 10.20) để đo khoảng cách giữa các cụm.

Khi quan sát kỹ Phương trình (10.19), chúng ta thấy rằng việc hợp nhất hai cụm có thể không $P(C_{j1} \cup C_{j2})$

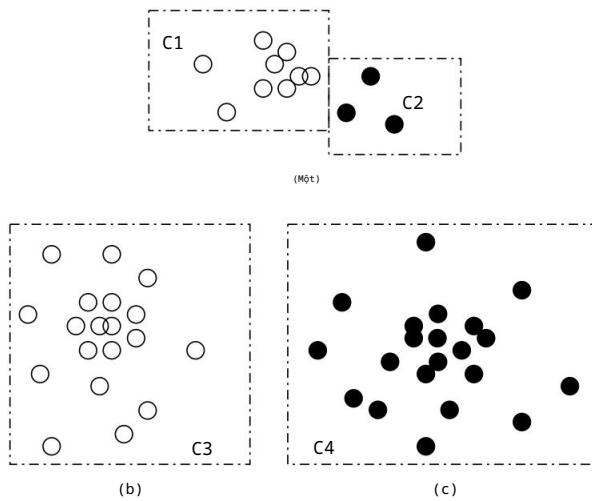
luôn dẫn đến sự cải thiện về chất lượng cụm, nghĩa là, $P(C_{j1})P(C_{j2})$ có thể nhỏ hơn 1. Ví dụ, giả sử rằng các hàm phân phối Gaussian được sử dụng trong mô hình của Hình 10.11. Mặc dù việc hợp nhất các cụm $C1$ và $C2$ tạo ra một cụm phù hợp hơn với phân phối Gaussian, việc hợp nhất các cụm $C3$ và $C4$ làm giảm chất lượng cụm vì không có hàm Gaussian nào có thể phù hợp tốt với cụm đã hợp nhất.

Dựa trên quan sát này, một sơ đồ phân cụ phân cấp xác suất có thể bắt đầu nếu khoảng cách giữa với một cụm cho mỗi đối tượng và hợp nhất hai cụm, C_i và C_j , chúng là số âm. Trong mỗi lần lặp, chúng tôi cố gắng tìm C_i và C_j để tối đa hóa $\log \frac{P(C_i \cup C_j)}{P(C_i)P(C_j)}$. Lặp tiếp tục miễn là $\log > 0$, nghĩa là ~~miễn~~ có sự cải thiện về chất lượng cụm. Mã giả được đưa ra trong Hình 10.12.

Các phương pháp phân cụ phân cấp xác suất dễ hiểu và nhìn chung có cùng hiệu quả như các phương pháp phân cụ phân cấp kết tụ thuật toán; trên thực tế, chúng chia sẻ cùng một khuôn khổ. Các mô hình xác suất dễ diễn giải hơn, nhưng đôi khi kém linh hoạt hơn các số liệu khoảng cách. Các mô hình xác suất có thể xử lý dữ liệu được quan sát một phần. Ví dụ, với một tập dữ liệu đa chiều trong đó một số đối tượng có giá trị bị thiếu trên một số chiều, chúng ta có thể học một mô hình Gaussian trên mỗi chiều một cách độc lập bằng cách sử dụng các giá trị được quan sát trên chiều đó. Phân cụ phân cấp cụm kết quả đạt được mục tiêu tối ưu hóa là khớp dữ liệu với các mô hình xác suất đã chọn.

Một nhược điểm của việc sử dụng phân cụm phân cấp xác suất là nó chỉ đưa ra một phân cấp đối với một mô hình xác suất đã chọn. Nó không thể xử lý được sự không chắc chắn của các phân cấp cụm. Với một tập dữ liệu cho trước, có thể tồn tại nhiều phân cấp

470 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản



Hình 10.11 Việc hợp nhất các cụm trong phân cấp xác suất: (a) Việc hợp nhất các cụm C1 và C2 dẫn đến sự gia tăng chất lượng cụm tổng thể, nhưng việc hợp nhất các cụm (b) C3 và (c) C4 thì không.

Thuật toán: Thuật toán phân cụ phân cấp xác suất.

Đầu vào:

$D = \{o_1, \dots, o_n\}$: một tập dữ liệu chứa n đối tượng;

Đầu ra: Một hệ thống phân cấp các cụm.

Phương pháp:

(1) tạo một cụm cho mỗi đối tượng $C_i = \{o_i\}$, $1 \leq i \leq n$; (2) cho $i = 1$ đến n (3)

(4) tìm cặp cụm C_i và C_j sao cho $C_i, C_j = \operatorname{argmax}_{i=j} \log \frac{P(C_i | C_j)}{P(C_i)P(C_j)}$;
nếu $\log \frac{P(C_i | C_j)}{P(C_i)P(C_j)}$ hợp nhất C_i và C_j ;
(5) nếu không thi dừng;

Hình 10.12 Thuật toán phân cụ phân cấp xác suất.

phù hợp với dữ liệu quan sát được. Cả phương pháp tiếp cận thuật toán và phương pháp tiếp cận xác suất đều không thể tìm ra phân phối của các hệ thống phân cấp như vậy. Gần đây, các mô hình cấu trúc cây Bayesian đã được phát triển để xử lý các vấn đề như vậy. Bayesian và các phương pháp phân cụm xác suất tinh vi khác được coi là các chủ đề nâng cao và không được đề cập trong cuốn sách này.

10.4 Phương pháp dựa trên mật độ

Các phương pháp phân vùng và phân cấp được thiết kế để tìm các cụm hình cầu.

Họ gặp khó khăn trong việc tìm các cụm có hình dạng tùy ý như hình chữ "S" và các cụm hình bầu dục trong Hình 10.13. Với dữ liệu như vậy, họ có thể sẽ xác định không chính xác các vùng lồi, nơi nhiều hoặc các giá trị ngoại lai được bao gồm trong các cụm.

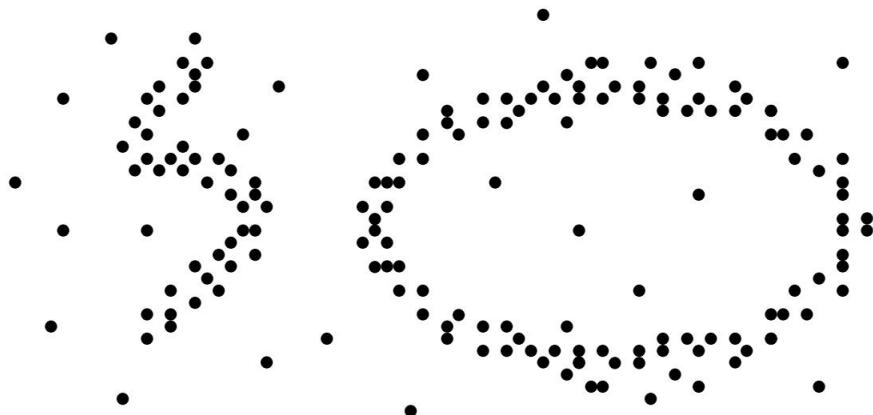
Để tìm các cụm có hình dạng tùy ý, chúng ta có thể mô hình hóa các cụm như các vùng dày đặc trong không gian dữ liệu, được phân tách bằng các vùng thưa thớt. Đây là chiến lược chính đáng sau các phương pháp phân cụm dựa trên mật độ, có thể khám phá các cụm có hình dạng không phải hình cầu. Trong phần này, bạn sẽ tìm hiểu các kỹ thuật cơ bản của phân cụm dựa trên mật độ bằng cách nghiên cứu ba phương pháp tiêu biểu, cụ thể là DBSCAN (Phần 10.4.1), OPTICS (Phần 10.4.2) và DENCLUE (Phần 10.4.3).

10.4.1 DBSCAN: Phân cụm dựa trên mật độ dựa trên các vùng được kết nối với mật độ cao

"Làm thế nào chúng ta có thể tìm thấy các vùng dày đặc trong phân cụm dựa trên mật độ?" Mật độ của một đối tượng o có thể được đo bằng số lượng các đối tượng gần với o . DBSCAN (Phân cụm không gian dựa trên mật độ của các ứng dụng có nhiều) tìm thấy các đối tượng cốt lõi, tức là các đối tượng có vùng lân cận dày đặc. Nó kết nối các đối tượng cốt lõi và vùng lân cận của chúng để tạo thành các vùng dày đặc dưới dạng các cụm.

"DBSCAN định lượng vùng lân cận của một đối tượng như thế nào?" Tham số do người dùng chỉ định $\epsilon > 0$ được sử dụng để chỉ định bán kính của vùng lân cận mà chúng tôi xem xét cho mọi đối tượng. -Lân cận của một vật thể o là không gian trong bán kính có tâm tại o .

Do kích thước vùng lân cận cố định được tham số hóa bởi , mật độ của một vùng lân cận có thể được đo đơn giản bằng số lượng đối tượng trong vùng lân cận. Để xác định xem một vùng lân cận có dày đặc hay không, DBSCAN sử dụng một



Hình 10.13 Các cụm có hình dạng tùy ý.

tham số, MinPts, chỉ định ngưỡng mật độ của các vùng dày đặc. Một đối tượng là đối tượng cốt lõi nếu -vùng lân cận của đối tượng chứa ít nhất MinPts đối tượng. Đối tượng cốt lõi là trụ cột của các vùng dày đặc.

Với một tập hợp D đối tượng, chúng ta có thể xác định tất cả các đối tượng cốt lõi liên quan đến các tham số đã cho và MinPts. Nhiệm vụ phân cụm được giảm xuống bằng cách sử dụng các đối tượng cốt lõi và vùng lân cận của chúng để tạo thành các vùng dày đặc, trong đó các vùng dày đặc là các cụm. Đối với một đối tượng lõi q và một đối tượng p, chúng ta nói rằng p có thể tiếp cận trực tiếp mật độ từ q (đối với và MinPts) nếu p nằm trong -lân cận của q. Rõ ràng, một đối tượng p có thể tiếp cận trực tiếp mật độ từ một đối tượng q khác nếu và chỉ nếu q là một đối tượng lõi và p nằm trong -lân cận của q. Sử dụng mối quan hệ có thể tiếp cận trực tiếp mật độ, một đối tượng lõi có thể "mang" tất cả các đối tượng từ -lân cận của nó vào một vùng dày đặc.

"Làm thế nào chúng ta có thể lắp ráp một vùng dày đặc lớn bằng cách sử dụng các vùng dày đặc nhỏ có tâm là các đối tượng lõi?" Trong DBSCAN, p có thể tiếp cận được bằng mật độ từ q (đối với và MinPts trong D) nếu có một chuỗi các đối tượng p_1, \dots, p_n , sao cho $p_1 = q$, $p_n = p$ và p_{i+1} có thể tiếp cận được bằng mật độ trực tiếp từ p_i đối với và MinPts, với $1 \leq i \leq n$, $p_i \in D$. Lưu ý rằng khả năng tiếp cận được bằng mật độ không phải là một quan hệ tương đương vì nó không đối xứng. Nếu cả o_1 và o_2 đều là các đối tượng lõi và o_1 có thể tiếp cận được bằng mật độ từ o_2 , thì o_2 có thể tiếp cận được bằng mật độ từ o_1 . Tuy nhiên, nếu o_2 là một đối tượng lõi nhưng o_1 thì không, thì o_1 có thể tiếp cận được bằng mật độ từ o_2 , nhưng không ngược lại.

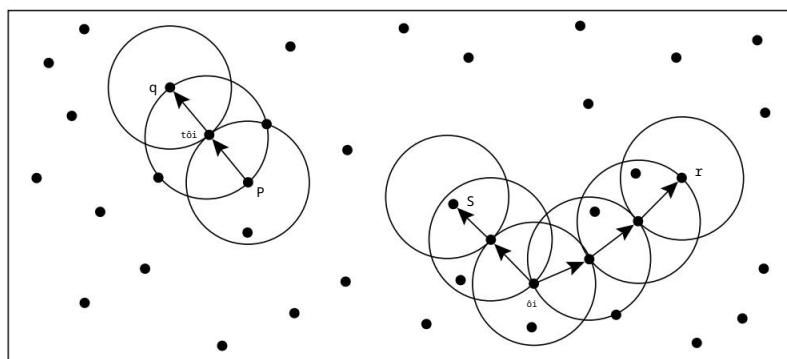
Để kết nối các đối tượng lõi cũng như các đối tượng lân cận của chúng trong một vùng dày đặc, DBSCAN sử dụng khái niệm về mật độ kết nối. Hai đối tượng $p_1, p_2 \in D$ được kết nối mật độ đối với và MinPts nếu có một đối tượng $q \in D$ sao cho cả p_1 và p_2 đều có thể tiếp cận được mật độ từ q đối với và MinPts. Không giống như khả năng tiếp cận được mật độ, mật độ kết nối là một quan hệ tương đương. Có thể dễ dàng chứng minh rằng, đối với các đối tượng o_1, o_2 và o_3 , nếu o_1 và o_2 được kết nối mật độ, và o_2 và o_3 được kết nối mật độ, thì o_1 và o_3 cũng vậy . ■

Ví dụ 10.7 Khả năng tiếp cận mật độ và khả năng kết nối mật độ. Xem xét Hình 10.14 cho một được biểu diễn bằng bán kính của các vòng tròn và, giả sử, đặt MinPts = 3.

Trong số các điểm được gắn nhãn, m , p , o , r là các đối tượng cốt lõi vì mỗi đối tượng nằm trong một -lân cận chứa ít nhất ba điểm. Đối tượng q có thể tiếp cận trực tiếp mật độ từ m . Đối tượng m có thể tiếp cận trực tiếp mật độ từ p và ngược lại.

Đối tượng q có thể tiếp cận mật độ (gián tiếp) từ p vì q có thể tiếp cận mật độ trực tiếp từ m và m có thể tiếp cận mật độ trực tiếp từ p . Tuy nhiên, p không thể tiếp cận mật độ từ q vì q không phải là đối tượng cốt lõi. Tương tự như vậy, r và s có thể tiếp cận mật độ từ o và o có thể tiếp cận mật độ từ r . Do đó, o , r và s đều có liên kết mật độ. ■

Ta có thể sử dụng phép đóng của mật độ-liên thông để tìm các vùng dày đặc liên thông dưới dạng cụm. Mỗi tập đóng là một cụm dựa trên mật độ. Một tập con $C \subseteq D$ là một cụm nếu (1) đối với bất kỳ hai đối tượng $o_1, o_2 \in C$, o_1 và o_2 là liên thông mật độ; và (2) không tồn tại một đối tượng $o \in C$ và một đối tượng khác $o' \in (D \setminus C)$ sao cho o và o' là liên thông mật độ.



Hình 10.14 Khả năng tiếp cận mảng độ và kết nối mảng độ trong cụm dựa trên mảng độ. Nguồn: Dựa trên Ester, Kriegel, Sander và Xu [EKSX96].

"DBSCAN tìm cụm như thế nào?" Ban đầu, tất cả các đối tượng trong tập dữ liệu D nhất định được đánh dấu là "chưa được truy cập". DBSCAN chọn ngẫu nhiên một đối tượng p chưa được truy cập, đánh dấu p là "đã được truy cập" và kiểm tra xem vùng lân cận của p có chứa ít nhất MinPts đối tượng hay không. Nếu không, p được đánh dấu là điểm nhiễu. Nếu không, một cụm C mới được tạo cho p và tất cả các đối tượng trong -neighborhood của p được thêm vào một tập ứng viên, N. DBSCAN lặp đi lặp lại thêm vào C các đối tượng trong N không thuộc bất kỳ cụm nào. Trong quá trình này, đối với một đối tượng p trong N mang nhãn "chưa ghé thăm", DBSCAN đánh dấu đối tượng đó là "đã ghé thăm" và kiểm tra -neighborhood của đối tượng đó. Nếu -neighborhood của p có ít nhất MinPts đối tượng, các đối tượng trong -neighborhood của p được thêm vào N. DBSCAN tiếp tục thêm các đối tượng vào C cho đến khi C không mở rộng được nữa, tức là N trống. Tại thời điểm này, cụm C đã hoàn thành và do đó được đưa ra.

Để tìm cụm tiếp theo, DBSCAN chọn ngẫu nhiên một đối tượng chưa được thăm từ các đối tượng còn lại. Quá trình phân cụm tiếp tục cho đến khi tất cả các đối tượng được thăm. Mã giả của thuật toán DBSCAN được đưa ra trong Hình 10.15.

Nếu chỉ số không gian được sử dụng, độ phức tạp tính toán của DBSCAN là $O(n \log n)$. Với n là số lượng đối tượng cơ sở dữ liệu. Nếu không, độ phức tạp là $O(n^2)$ thiết lập thích hợp của các tham số do người dùng xác định và MinPts, thuật toán có hiệu quả trong việc tìm các cụm có hình dạng tùy ý.

10.4.2 QUANG HỌC: Sắp xếp các điểm để xác định cấu trúc cụm

Mặc dù DBSCAN có thể nhóm các đối tượng với các tham số đầu vào như (bán kính tối đa của một vùng lân cận) và MinPts (số điểm tối thiểu cần thiết trong vùng lân cận của một đối tượng cốt lõi), nhưng nó lại làm cản trở người dùng với trách nhiệm lựa chọn các giá trị tham số sẽ dẫn đến việc khám phá ra các cụm có thể chấp nhận được. Đây là một vấn đề liên quan đến nhiều thuật toán nhóm khác. Các thiết lập tham số như vậy

474 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản

Thuật toán: DBSCAN: thuật toán phân cụm dựa trên mật độ.

Đầu vào:

- D: một tập dữ liệu chứa n đối tượng,
- : tham số bán kính và
- MinPts: ngưỡng mật độ khu vực lân cận.

Đầu ra: Một tập hợp các cụm dựa trên mật độ.

Phương pháp:

- (1) đánh dấu tất cả các đối tượng là chưa được truy cập;
- (2) làm
 - (3) chọn ngẫu nhiên một đối tượng chưa được truy cập p;
 - (4) đánh dấu p là đã truy cập;
 - (5) nếu -lân cận của p có ít nhất MinPts đối tượng
 - (6) tạo cụm C mới và thêm p vào C;
 - (7) cho N là tập hợp các đối tượng trong lân cận của p;
 - (8) cho mỗi điểm p trong N
 - (9) nếu p chưa được thăm
 - (10) đánh dấu p là đã truy cập;
 - (11) nếu vùng lân cận của p có ít nhất MinPts điểm,
 - (12) thêm những điểm đó vào N;
 - (12) nếu p chưa phải là thành viên của bất kỳ cụm nào, thêm p vào C;
 - (13) kết thúc cho
 - (14) đầu ra C;
 - (15) nếu không thi đánh dấu p là tiếng ồn;
 - (16) cho đến khi không còn đối tượng nào chưa được ghé thăm;

Hình 10.15 Thuật toán DBSCAN.

thường được thiết lập theo kinh nghiệm và khó xác định, đặc biệt là đối với các tập dữ liệu thực tế, có nhiều chiều. Hầu hết các thuật toán đều nhạy cảm với các giá trị tham số này: Hơi các thiết lập khác nhau có thể dẫn đến các cụm dữ liệu rất khác nhau. Hơn nữa, trong thế giới thực, các tập dữ liệu có chiều cao thường có phân phối rất lệch sao cho cấu trúc cụm nội tại của chúng có thể không được mô tả tốt bằng một tập hợp mật độ toàn cầu duy nhất tham số.

Lưu ý rằng các cụm dựa trên mật độ là đơn điệu đối với vùng lân cận ngưỡng. Nghĩa là, trong DBSCAN, đối với giá trị MinPts cố định và hai ngưỡng lân cận, $1 < 2$, một cụm C đối với 1 và MinPts phải là một tập hợp con của một cụm C đối với 2 và MinPts. Điều này có nghĩa là nếu hai đối tượng nằm trong mật độ dựa trên cụm, chúng cũng phải nằm trong cụm có yêu cầu mật độ thấp hơn.

Để khắc phục khó khăn trong việc sử dụng một tập hợp các tham số toàn cục trong phân tích cụm, một phương pháp phân tích cụm được gọi là OPTICS đã được đề xuất. OPTICS không rõ ràng tạo ra một cụm tập dữ liệu. Thay vào đó, nó đưa ra một cụm sắp xếp. Đây là một danh sách tuyến tính

của tất cả các đối tượng đang được phân tích và biểu diễn cấu trúc cụm dựa trên mảng độ của dữ liệu. Các đối tượng trong cụm dày đặc hơn được liệt kê gần nhau hơn trong thứ tự cụm. Thứ tự này tương đương với phân cụm dựa trên mảng độ thu được từ nhiều thiết lập tham số. Do đó, OPTICS không yêu cầu người dùng cung cấp mảng độ cụ thể. Thứ tự phân cụm có thể được sử dụng để trích xuất thông tin phân cụm cơ bản (ví dụ: trung tâm cụm hoặc cụm có hình dạng tùy ý), suy ra cấu trúc phân cụm nội tại cũng như cung cấp hình ảnh trực quan về phân cụm.

Để xây dựng các cụm khác nhau cùng lúc, các đối tượng được xử lý theo một thứ tự cụ thể. Thứ tự này chọn một đối tượng có thể đạt được mảng độ liên quan đến giá trị thấp nhất để các cụm có mảng độ cao hơn (thấp hơn) sẽ được hoàn thành trước. Dựa trên ý tưởng này, OPTICS cần hai thông tin quan trọng cho mỗi đối tượng:

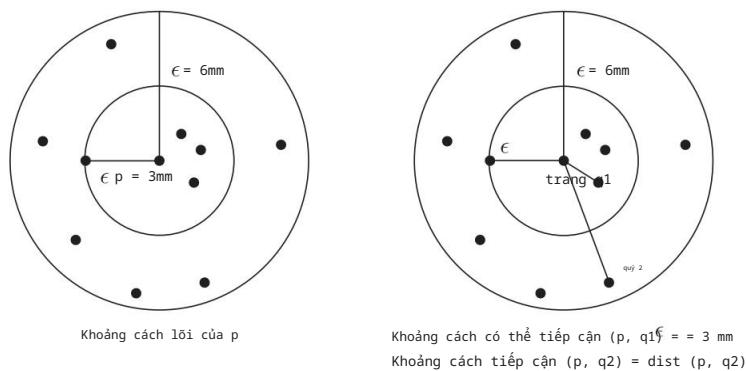
- Khoảng cách lõi của một đối tượng p là giá trị nhỏ nhất sao cho vùng lân cận của p có ít nhất MinPts đối tượng. Nghĩa là, là khoảng cách tối thiểu ngưỡng tách khiến p trở thành đối tượng cốt lõi. Nếu p không phải là đối tượng cốt lõi đối với và MinPts , khoảng cách cốt lõi của p không xác định.
- Khoảng cách có thể tiếp cận đến đối tượng p từ q là giá trị bán kính nhỏ nhất khiến p có thể tiếp cận được theo mảng độ từ q . Theo định nghĩa về khả năng tiếp cận theo mảng độ, q phải là một đối tượng lõi và p phải nằm trong vùng lân cận của q . Do đó, khoảng cách có thể tiếp cận từ q đến p là $\max(\text{core-distance}(q), \text{dist}(p, q))$. Nếu q không phải là một đối tượng lõi đối với và MinPts , thì khoảng cách có thể tiếp cận đến p từ q là không xác định.

Một đối tượng p có thể được tiếp cận trực tiếp từ nhiều đối tượng lõi. Do đó, p có thể có nhiều khoảng cách tiếp cận đối với các đối tượng lõi khác nhau. Khoảng cách tiếp cận nhỏ nhất của p đặc biệt được quan tâm vì nó cung cấp đường dẫn ngắn nhất mà p được kết nối với một cụm dày đặc.

Ví dụ 10.8 Khoảng cách lõi và khoảng cách khả năng tiếp cận. Hình 10.16 minh họa các khái niệm về khoảng cách lõi và khoảng cách khả năng tiếp cận. Giả sử rằng $= 6\text{ mm}$ và $\text{MinPts} = 5$. Khoảng cách lõi của p là khoảng cách, , giữa p và đối tượng dữ liệu gần thứ tư từ p . Khoảng cách có thể tiếp cận của q_1 từ p là khoảng cách lõi của p (tức là $= 3\text{mm}$) vì khoảng cách này lớn hơn khoảng cách Euclid từ p đến q_1 . Khoảng cách có thể tiếp cận của q_2 đối với p là khoảng cách Euclid từ p đến q_2 vì khoảng cách này lớn hơn khoảng cách lõi của p .

OPTICS tính toán thứ tự của tất cả các đối tượng trong một cơ sở dữ liệu nhất định và, đối với mỗi đối tượng trong cơ sở dữ liệu, lưu trữ khoảng cách lõi và khoảng cách tiếp cận phù hợp. OPTICS duy trì một danh sách có tên là OrderSeeds để tạo thứ tự đầu ra. Các đối tượng trong Order-Seeds được sắp xếp theo khoảng cách tiếp cận từ các đối tượng lõi gần nhất tương ứng của chúng, tức là theo khoảng cách tiếp cận nhỏ nhất của mỗi đối tượng.

OPTICS bắt đầu với một đối tượng tùy ý từ cơ sở dữ liệu đầu vào là đối tượng hiện tại, p . Nó lấy -neighborhood của p , xác định core-distance và đặt reachability-distance thành undefined. Đối tượng hiện tại, p , sau đó được ghi vào đầu ra.



Hình 10.16 Thuật ngữ QUANG HỌC. Nguồn: Dựa trên Ankerst, Breunig, Kriegel và Sander [ABKS99].

Nếu p không phải là đối tượng cốt lõi, OPTICS chỉ cần chuyển sang đối tượng tiếp theo trong danh sách OrderSeeds (hoặc cơ sở dữ liệu đầu vào nếu OrderSeeds trống). Nếu p là đối tượng cốt lõi, thì đối với mỗi đối tượng, q, trong -neighborhood của p, OPTICS cập nhật reachability-distance của nó từ p và chèn q vào OrderSeeds nếu q vẫn chưa được xử lý. Lặp lại tiếp tục cho đến khi đầu vào được sử dụng hoàn toàn và OrderSeeds trống.

Thứ tự cụm của một tập dữ liệu có thể được biểu diễn bằng đồ họa, giúp trực quan hóa và hiểu cấu trúc cụm trong một tập dữ liệu. Ví dụ, Hình 10.17 là biểu đồ khả năng tiếp cận cho một tập dữ liệu 2 chiều đơn giản, biểu diễn tổng quan chung về cách dữ liệu được cấu trúc và cụm. Các đối tượng dữ liệu được biểu diễn theo thứ tự cụm (trục ngang) cùng với khoảng cách khả năng tiếp cận tương ứng của chúng (trục dọc). Ba "điểm lõi" Gaussian trong biểu đồ phản ánh ba cụm trong tập dữ liệu. Các phương pháp cũng đã được phát triển để xem các cấu trúc cụm của dữ liệu nhiều chiều ở nhiều cấp độ chi tiết khác nhau.

Cấu trúc của thuật toán OPTICS rất giống với cấu trúc của DBSCAN. Do đó, hai thuật toán có cùng độ phức tạp về thời gian. Độ phức tạp là $O(n \log n)$ nếu sử dụng chỉ số không gian và $O(n^2)$ nếu không, trong đó n là số đối tượng.

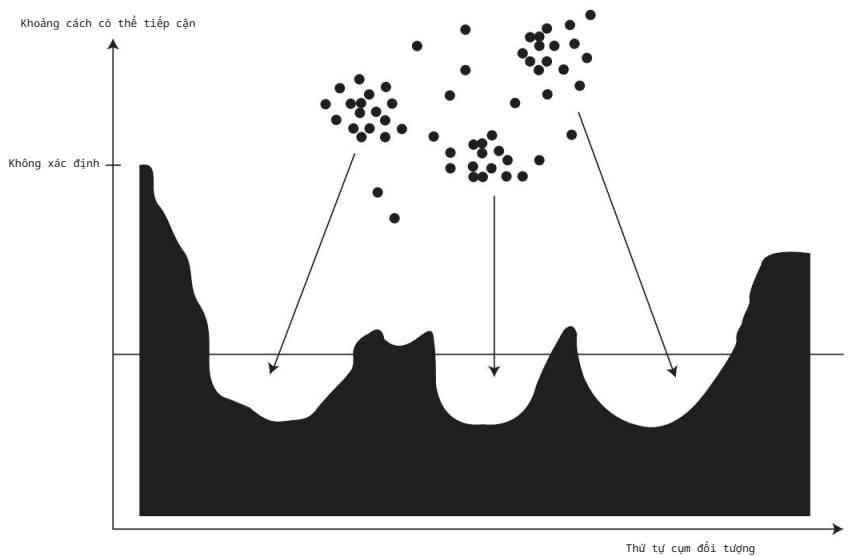
10.4.3 DENCLUE: Phân cụm dựa trên mật độ

Các hàm phân phối

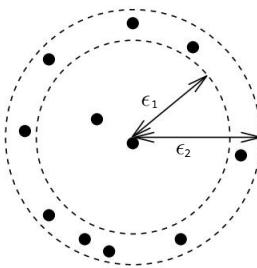
Ước tính mật độ là một vấn đề cốt lõi trong các phương pháp phân cụm dựa trên mật độ.

DENCLUE (DENsity-based CLUstEring) là một phương pháp phân cụm dựa trên một tập hợp các hàm phân phối mật độ. Trước tiên, chúng tôi cung cấp một số thông tin cơ bản về ước tính mật độ, sau đó mô tả thuật toán DENCLUE.

Trong xác suất và thống kê, ước tính mật độ là ước tính hàm mật độ xác suất cơ bản không thể quan sát được dựa trên một tập hợp dữ liệu quan sát được. Trong bối cảnh của cụm dựa trên mật độ, hàm mật độ xác suất cơ bản không thể quan sát được là phân phối thực sự của quần thể của tất cả các đối tượng có thể được phân tích. Tập dữ liệu quan sát được coi là một mẫu ngẫu nhiên từ quần thể đó.



Hình 10.17 Sắp xếp cụm trong OPTICS. Nguồn: Chuyển thể từ Ankerst, Breunig, Kriegel và Sander [ABKS99].



Hình 10.18 Sự tinh tế trong ước tính mật độ trong DBSCAN và OPTICS: Tăng nhẹ bán kính lân cận từ 2 đến 1 đến mật độ cao hơn nhiều. 1 đến

Trong DBSCAN và OPTICS, mật độ được tính bằng cách đếm số lượng vật thể trong một vùng lân cận được xác định bởi tham số bán kính. Các ước tính mật độ như vậy có thể rất nhạy cảm với giá trị bán kính được sử dụng. Ví dụ, trong Hình 10.18, mật độ thay đổi đáng kể khi bán kính tăng một lượng nhỏ.

Để khắc phục vấn đề này, có thể sử dụng ước tính mật độ hạt nhân, đây là phương pháp ước tính mật độ phi tham số từ thống kê. Ý tưởng chung đằng sau ước tính mật độ hạt nhân rất đơn giản. Chúng tôi coi một đối tượng được quan sát là một chỉ báo của

mật độ xác suất cao trong vùng xung quanh. Mật độ xác suất tại một điểm phụ thuộc vào khoảng cách từ điểm này đến các vật thể quan sát được.

Về mặt hình thức, hãy để x_1, \dots, x_N là một mẫu độc lập và phân phối giống hệt nhau của một biến ngẫu nhiên f . Xấp xỉ mật độ hạt nhân của hàm mật độ xác suất là

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right), \quad (10.21)$$

trong đó $K()$ là một hạt nhân và h là băng thông đóng vai trò là tham số làm mịn. Một hạt nhân có thể được coi là một hàm mô hình hóa ảnh hưởng của một điểm mẫu trong vùng lân cận của nó. Về mặt kỹ thuật, hạt nhân $K()$ là một hàm tích phân có giá trị thực không âm phải đáp ứng hai yêu cầu: $K(u)du = 1$ và $K(-u) = K(u)$ đối với mọi giá trị của u . Một hạt nhân thường được sử dụng là một hàm Gaussian chuẩn với giá trị trung bình là 0 và phương sai là 1:

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - x_i)^2}{2h^2}}. \quad (10.22)$$

DENCLUE sử dụng hạt nhân Gaussian để ước tính mật độ dựa trên tập hợp các đối tượng được nhóm lại. Một điểm x được gọi là điểm thu hút mật độ nếu nó là cực đại cục bộ của hàm mật độ ước tính. Để tránh các điểm cực đại cục bộ tầm thường, DENCLUE sử dụng ngưỡng nhiễu, ξ và chỉ xem xét các điểm thu hút mật độ x này. Các điểm thu hút mật độ không tầm sao cho $f(x) \geq \xi$. Thường này là tầm của các cụm.

Các đối tượng đang được phân tích được gán vào các cụm thông qua các điểm thu hút mật độ bằng cách sử dụng quy trình leo dồi từng bước. Đối với một đối tượng, x , quy trình leo dồi bắt đầu từ x và được hướng dẫn bởi độ dốc của hàm mật độ ước tính. Nghĩa là, điểm thu hút mật độ cho x được tính như

$$\hat{x}_n = x$$

$$\hat{x}_{j+1} = x_j + \frac{\hat{f}(x)}{\delta |\hat{f}'(x_j)|}, \quad (10.23)$$

trong đó δ là một tham số để kiểm soát tốc độ hội tụ và

$$\hat{f}(x) = \frac{1}{h^{d+2n} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)}. \quad (10.24)$$

Quy trình leo dồi dừng lại ở bước $k > 0$ nếu $f(x)$ mật độ hấp $\hat{f}(x) < \epsilon$, và gán x cho \hat{x} quy trình leo $= x$. Một đối tượng x là một giá trị ngoại lai hoặc nhiễu nếu nó hội tụ ở vùng dồi. Đến cuối $\hat{f}(x) < \xi$.

Một cụm trong DENCLUE là một tập hợp các điểm thu hút mật độ X và một tập hợp các đối tượng đầu vào C sao cho mỗi đối tượng trong C được gán cho một điểm thu hút mật độ trong X và tồn tại một đường dẫn giữa mọi cặp điểm thu hút mật độ trong đó mật độ lớn hơn ξ . Bằng cách sử dụng nhiều điểm thu hút mật độ được kết nối bằng các đường dẫn, DENCLUE có thể tìm thấy các cụm có hình dạng tùy ý.

DENCLUE có một số ưu điểm. Nó có thể được coi là sự tổng quát hóa của một số phương pháp phân cụm nổi tiếng như các phương pháp liên kết đơn và DBSCAN. Hơn nữa, DENCLUE không thay đổi so với nhiễu. Ước tính mật độ hạt nhân có thể giảm hiệu quả ảnh hưởng của nhiễu bằng cách phân phối nhiễu đồng đều vào dữ liệu đầu vào.

10.5 Phương pháp dựa trên lưới

Các phương pháp phân cụm được thảo luận cho đến nay đều dựa trên dữ liệu-chung phân vùng tập hợp các đối tượng và thích ứng với sự phân bố của các đối tượng trong không gian nhung. Ngoài ra, phương pháp phân cụm dựa trên lưới sử dụng phương pháp tiếp cận dựa trên không gian bằng cách phân vùng không gian nhung thành các ô độc lập với sự phân bố của các đối tượng đầu vào.

Phương pháp phân cụm dựa trên lưới sử dụng cấu trúc dữ liệu lưới đa độ phân giải. Nó lượng tử hóa không gian đối tượng thành một số lượng hữu hạn các ô tạo thành một cấu trúc lưới mà trên đó tất cả các hoạt động phân cụm được thực hiện. Ưu điểm chính của phương pháp này là thời gian xử lý nhanh, thường không phụ thuộc vào số lượng đối tượng dữ liệu, nhưng chỉ phụ thuộc vào số lượng ô trong mỗi chiều trong lượng tử hóa

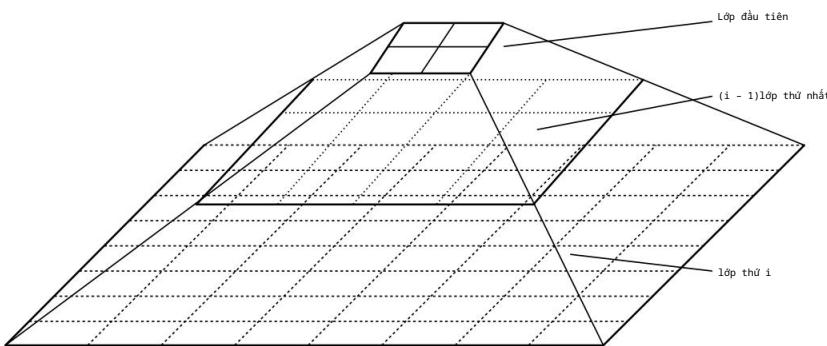
không gian.

Trong phần này, chúng tôi minh họa việc phân cụm dựa trên lưới bằng hai ví dụ điển hình. STING (Phần 10.5.1) khai phá thông tin thống kê được lưu trữ trong các ô lưới. CLIQUE (Phần 10.5.2) thể hiện một phương pháp tiếp cận dựa trên lưới và mật độ để phân cụm không gian con trong không gian dữ liệu có nhiều chiều.

10.5.1 STING: Lưới thông tin thống kê

STING là một kỹ thuật phân cụm đa độ phân giải dựa trên lưới trong đó vùng không gian nhung của các đối tượng đầu vào được chia thành các ô hình chữ nhật. Không gian có thể được chia theo cách phân cấp và đệ quy. Một số cấp độ của các ô hình chữ nhật như vậy tương ứng với các cấp độ phân giải khác nhau và tạo thành một cấu trúc phân cấp: Mỗi ô ở cấp độ cao được phân vùng để tạo thành một số ô ở cấp độ thấp hơn tiếp theo. Thông tin thống kê liên quan đến các thuộc tính trong mỗi ô lưới, chẳng hạn như giá trị trung bình, giá trị tối đa và giá trị tối thiểu, được tính toán trước và lưu trữ dưới dạng các tham số thống kê. Các tham số thống kê này hữu ích cho việc xử lý truy vấn và cho các tác vụ phân tích dữ liệu khác.

Hình 10.19 cho thấy cấu trúc phân cấp cho cụm STING. Các tham số thống kê của các ô cấp cao hơn có thể dễ dàng được tính toán từ các tham số của các ô cấp thấp hơn. Các tham số này bao gồm: tham số độc lập với thuộc tính, count; và các tham số phụ thuộc vào thuộc tính, mean, stdev (độ lệch chuẩn), min (tối thiểu), max (tối đa) và loại phân phối mà giá trị thuộc tính trong ô tuân theo như chuẩn, đồng đều, mũ hoặc không có (nếu phân phối không xác định). Ở đây, thuộc tính là một biện pháp được chọn để phân tích như giá cho các đối tượng nhà ở. Khi dữ liệu được tải vào cơ sở dữ liệu, các tham số count, mean, stdev, min và max của các ô cấp thấp nhất được tính toán trực tiếp từ dữ liệu. Giá trị phân phối có thể được người dùng chỉ định nếu loại phân phối được biết.



Hình 10.19 Cấu trúc phân cấp cho cụm STING.

trước đó hoặc thu được bằng các kiểm định giả thuyết như χ^2 kiểm tra. Loại phân phối của một ô cấp cao hơn có thể được tính toán dựa trên phần lớn các loại phân phối của các ô cấp thấp tương ứng của nó kết hợp với một quy trình lọc ngưỡng. Nếu các phân phối của các ô cấp thấp hơn không đồng nhất với nhau và không vượt qua được kiểm định ngưỡng, loại phân phối của ô cấp cao được đặt thành không.

"Thông tin thống kê nào hữu ích như thế nào cho việc trả lời truy vấn?" Các tham số thống kê có thể được sử dụng theo cách từ trên xuống, dựa trên lối như sau. Đầu tiên, một lớp trong cấu trúc phân cấp được xác định từ đó quá trình trả lời truy vấn sẽ bắt đầu. Lớp này thường chứa một số lượng nhỏ các ô. Đối với mỗi ô trong lớp hiện tại, chúng tôi tính toán khoảng tin cậy (hoặc phạm vi xác suất ước tính) phản ánh mức độ liên quan của ô với truy vấn đã cho. Các ô không liên quan sẽ bị loại khỏi quá trình xem xét tiếp theo. Quá trình xử lý ở cấp độ thấp hơn tiếp theo chỉ kiểm tra các ô liên quan còn lại. Quá trình này được lặp lại cho đến khi đạt đến lớp dưới cùng. Tại thời điểm này, nếu thông số kỹ thuật truy vấn được đáp ứng, các vùng của các ô liên quan thỏa mãn truy vấn sẽ được trả về. Nếu không, dữ liệu rời vào các ô liên quan sẽ được truy xuất và xử lý thêm cho đến khi chúng đáp ứng các yêu cầu của truy vấn.

Một đặc tính thú vị của STING là nó tiếp cận kết quả phân cụm của DBSCAN nếu độ chi tiết tiếp cận θ (tức là, hướng tới dữ liệu cấp rất thấp). Nói cách khác, sử dụng thông tin về số lượng và kích thước ô, các cụm dày đặc có thể được xác định gần đúng bằng STING. Do đó, STING cũng có thể được coi là phương pháp phân cụm dựa trên mật độ.

"STING cung cấp những lợi thế gì so với các phương pháp phân cụm khác?" STING cung cấp một số lợi thế: (1) tính toán dựa trên lối không phụ thuộc vào truy vấn vì thông tin thống kê được lưu trữ trong mỗi ô biểu diễn thông tin tóm tắt của dữ liệu trong ô lối, không phụ thuộc vào truy vấn; (2) cấu trúc lối tạo điều kiện thuận lợi cho xử lý song song và cập nhật giá tăng; và (3) hiệu quả của phương pháp là một lợi thế lớn: STING duyệt qua cơ sở dữ liệu một lần để tính toán các tham số thống kê của các ô và do đó độ phức tạp về thời gian của việc tạo cụm là $O(n)$, trong đó n là tổng số đối tượng. Sau khi tạo cấu trúc phân cấp, thời gian xử lý truy vấn

là 0(g), trong đó g là tổng số ô lưới ở mức thấp nhất, thường nhỏ hơn nhiều so với n.

Vì STING sử dụng phương pháp phân tích cụm đa độ phân giải, nên chất lượng của cụm STING phụ thuộc vào độ chi tiết của mức thấp nhất của cấu trúc lưới. Nếu độ chi tiết rất mịn, chi phí xử lý sẽ tăng đáng kể; tuy nhiên, nếu mức dưới cùng của cấu trúc lưới quá thô, nó có thể làm giảm chất lượng của phân tích cụm. Hơn nữa, STING không xem xét mối quan hệ không gian giữa các con và các ô lân cận của chúng để xây dựng một ô cha. Do đó, hình dạng của các cụm kết quả là thẳng hướng, nghĩ a là tất cả các ranh giới cụm đều nằm ngang hoặc dọc và không phát hiện ra ranh giới chéo. Điều này có thể làm giảm chất lượng và độ chính xác của các cụm mặc dù thời gian xử lý của kỹ thuật này nhanh.

10.5.2 CLIQUE: Một phương pháp phân cụm không gian con giống Apriori

Một đối tượng dữ liệu thường có hàng chục thuộc tính, nhiều trong số đó có thể không liên quan. Giá trị của các thuộc tính có thể thay đổi đáng kể. Các yếu tố này có thể khiến việc xác định các cụm trải dài toàn bộ không gian dữ liệu trở nên khó khăn. Thay vào đó, có thể có ý nghĩa a hơn khi tìm kiếm các cụm trong các không gian con khác nhau của dữ liệu. Ví dụ, hãy xem xét một ứng dụng tin học y tế trong đó hồ sơ bệnh nhân chứa các thuộc tính mở rộng mô tả thông tin cá nhân, nhiều triệu chứng, tình trạng và tiền sử gia đình.

Việc tìm ra một nhóm bệnh nhân tiềm thường mà tất cả hoặc thậm chí hầu hết các thuộc tính đều đồng ý mạnh mẽ là điều không thể. Ví dụ, ở những bệnh nhân có gia cầm, độ tuổi, giới tính và các thuộc tính công việc có thể thay đổi đáng kể trong một phạm vi giá trị rộng. Do đó, có thể khó tìm thấy một cụm như vậy trong toàn bộ không gian dữ liệu. Thay vào đó, bằng cách tìm kiếm trong các không gian con, chúng ta có thể tìm thấy một cụm bệnh nhân tương tự trong không gian có chiều thấp hơn (ví dụ: những bệnh nhân tương tự nhau về các triệu chứng như sốt cao, ho nhưng không sổ mũi và ở độ tuổi từ 3 đến 16).

CLIQUE (Clustering In QUEST) là một phương pháp dựa trên lưới đơn giản để tìm các cụm dựa trên mật độ trong các không gian con. CLIQUE phân vùng mỗi chiều thành các khoảng không chồng lấn, do đó phân vùng toàn bộ không gian nhúng của các đối tượng dữ liệu thành các ô.

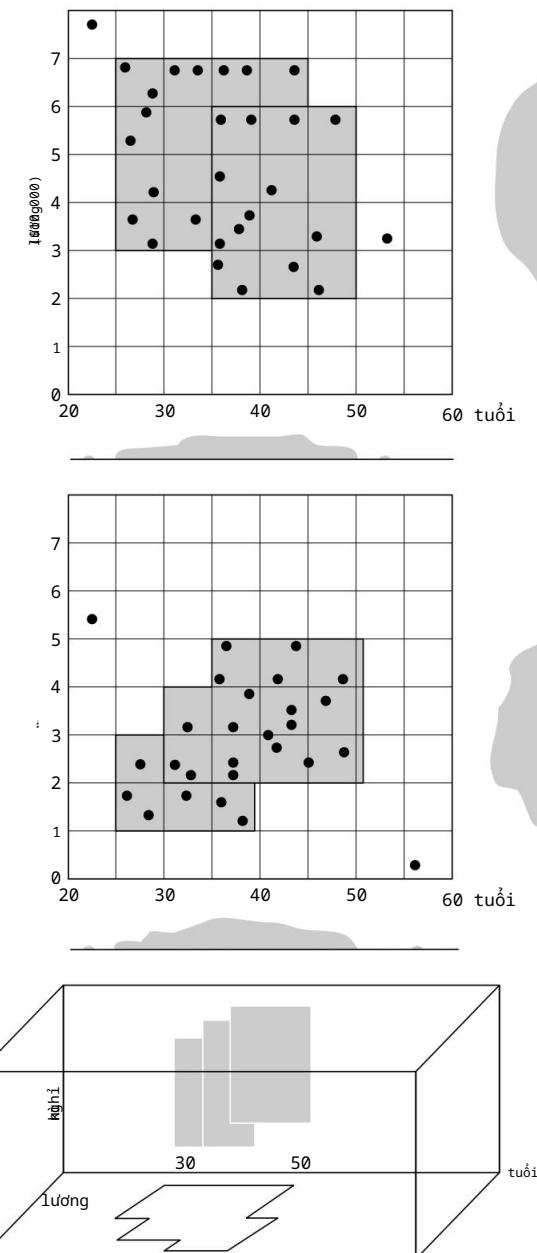
Nó sử dụng ngưỡng mật độ để xác định các ô dày đặc và các ô thưa thớt. Một ô được coi là dày đặc nếu số lượng đối tượng được ánh xạ tới nó vượt quá ngưỡng mật độ.

Chiến lược chính đằng sau CLIQUE để xác định không gian tìm kiếm ứng viên sử dụng tính đơn điệu của các ô dày đặc liên quan đến tính đa chiều. Điều này dựa trên tính chất Apriori được sử dụng trong khai thác mẫu thường xuyên và quy tắc liên kết (Chương 6). Trong bối cảnh của các cụm trong không gian con, tính đơn điệu nói lên điều sau. Một ô k chiều c ($k > 1$) chỉ có thể có ít nhất 1 điểm nếu mọi phép chiếu ($k - 1$) chiều của c, là một ô trong không gian con ($k - 1$) chiều, có ít nhất 1 điểm. Hãy xem xét Hình 10.20, trong đó không gian dữ liệu nhúng chứa ba chiều: tuổi, lương và kỳ nghỉ.

Một ô 2 chiều, chẳng hạn trong không gian con được hình thành bởi tuổi và lương, chỉ chứa 1 điểm nếu phép chiếu của ô này trong mọi chiều, tức là tuổi và lương, chứa ít nhất 1 điểm.

CLIQUE thực hiện phân cụm theo hai bước. Ở bước đầu tiên, CLIQUE phân vùng không gian dữ liệu d chiều thành các đơn vị hình chữ nhật không chồng lấn, xác định các đơn vị dày đặc trong số này. CLIQUE tìm các ô dày đặc trong tất cả các không gian con. Để thực hiện như vậy,

482 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản



Hình 10.20 Các đơn vị dày đặc được tìm thấy liên quan đến độ tuổi cho các chiều lương và kỳ nghỉ được giao nhau để cung cấp không gian tìm kiếm ứng viên cho các đơn vị dày đặc có nhiều chiều hơn.

CLIQUE phân vùng mọi chiều thành các khoảng và xác định các khoảng chứa ít nhất 1 điểm, trong đó 1 là ngưỡng mật độ. Sau đó, CLIQUE lặp lại nối hai ô dày đặc k chiều, c1 và c2, trong các không gian con (D_{i1}, \dots, D_{ik}) và (D_{j1}, \dots, D_{jk}), tương ứng, nếu $D_{i1} = D_{j1}, \dots, D_{ik} = D_{jk}$ và c_1 và c_2 chia sẻ cùng các khoảng trong các chiều đó. Hoạt động nối tạo ra một ô ứng viên c ($k + 1$) chiều mới trong không gian ($D_{i1}, \dots, D_{ik}, D_{jk}$). CLIQUE kiểm tra xem số điểm trong c có vượt qua ngưỡng mật độ hay không. Lặp lại kết thúc khi không thể tạo ra ứng viên nào hoặc không có ô ứng viên nào dày đặc.

Ở bước thứ hai, CLIQUE sử dụng các ô dày đặc trong mỗi không gian con để lắp ráp các cụm, có thể có hình dạng tùy ý. Ý tưởng là áp dụng nguyên lý Chiều dài mô tả tối thiểu (MDL) (Chương 8) để sử dụng các vùng cực đại để bao phủ các ô dày đặc được kết nối, trong đó một vùng cực đại là một siêu hình chữ nhật mà mọi ô rời vào vùng này đều dày đặc và vùng không thể mở rộng thêm theo bất kỳ chiều nào trong không gian con.

Việc tìm ra mô tả tốt nhất về một cụm nói chung là NP-Hard. Do đó, CLIQUE áp dụng một cách tiếp cận tham lam đơn giản. Nó bắt đầu với một ô dày đặc tùy ý, tìm một vùng tối đa bao phủ ô đó, sau đó xử lý các ô dày đặc còn lại chưa được bao phủ. Phương pháp tham lam kết thúc khi tất cả các ô dày đặc được bao phủ.

"CLIQUE hiệu quả như thế nào?" CLIQUE tự động tìm các không gian con có chiều cao nhất sao cho các cụm mật độ cao tồn tại trong các không gian con đó. Nó không nhạy cảm với thứ tự của các đối tượng đầu vào và không giả định bất kỳ phân phối dữ liệu chuẩn nào. Nó mở rộng tuyến tính theo kích thước của đầu vào và có khả năng mở rộng tốt khi số chiều trong dữ liệu tăng lên. Tuy nhiên, việc thu được một cụm có ý nghĩa phụ thuộc vào việc điều chỉnh đúng kích thước lưỡi (là một cấu trúc ổn định ở đây) và ngưỡng mật độ. Điều này có thể khó khăn trong thực tế vì kích thước lưỡi và ngưỡng mật độ được sử dụng trên tất cả các kết hợp chiều trong tập dữ liệu. Do đó, độ chính xác của kết quả cụm có thể bị giảm do tính đơn giản của phương pháp. Hơn nữa, đối với một vùng dày đặc nhất định, tất cả các phép chiếu của vùng đó lên các không gian con có chiều thấp hơn cũng sẽ dày đặc. Điều này có thể dẫn đến sự chồng chéo lớn giữa các vùng dày đặc được báo cáo. Hơn nữa, rất khó để tìm thấy các cụm có mật độ khá khác nhau trong các không gian con có chiều khác nhau.

Một số phần mở rộng cho cách tiếp cận này tuân theo một triết lý tương tự. Ví dụ, chúng ta có thể nghĩ về lưỡi như một tập hợp các thùng cố định. Thay vì sử dụng các thùng cố định cho mỗi chiều, chúng ta có thể sử dụng một chiến lược thích ứng, dựa trên dữ liệu để xác định động các thùng cho mỗi chiều dựa trên thống kê phân phối dữ liệu. Ngoài ra, thay vì sử dụng ngưỡng mật độ, chúng ta có thể sử dụng entropy (Chương 8) làm thước đo chất lượng của các cụm không gian con.

10.6 Đánh giá cụm

Bây giờ bạn đã biết cụm là gì và biết một số phương pháp cụm phổ biến. Bạn có thể hỏi, "Khi tôi thử một phương pháp cụm trên một tập dữ liệu, làm thế nào tôi có thể đánh giá xem kết quả cụm có tốt không?" Nói chung, đánh giá cụm đánh giá

tính khả thi của phân tích cụm trên một tập dữ liệu và chất lượng của kết quả do phương pháp cụm tạo ra. Các nhiệm vụ chính của đánh giá cụm bao gồm:

- **Đánh giá xu hướng phân cụm.** Trong nhiệm vụ này, đối với một tập dữ liệu nhất định, chúng tôi đánh giá xem có tồn tại cấu trúc phi ngẫu nhiên trong dữ liệu hay không. Áp dụng mù quáng phương pháp phân cụm trên một tập dữ liệu sẽ trả về các cụm; tuy nhiên, các cụm được khai thác có thể gây hiểu lầm. Phân tích phân cụm trên một tập dữ liệu chỉ có ý nghĩa khi có cấu trúc phi ngẫu nhiên trong dữ liệu.
- **Xác định số cụm trong một tập dữ liệu.** Một số thuật toán, chẳng hạn như k-means, yêu cầu số cụm trong một tập dữ liệu làm tham số. Hơn nữa, số cụm có thể được coi là một thống kê tóm tắt thứ vị và quan trọng của một tập dữ liệu. Do đó, cần ước tính số này ngay cả trước khi sử dụng thuật toán phân cụm để suy ra các cụm chi tiết.
- **Đo lường chất lượng cụm.** Sau khi áp dụng phương pháp cụm trên một tập dữ liệu, chúng tôi muốn đánh giá mức độ tốt của các cụm kết quả. Có thể sử dụng một số biện pháp. Một số phương pháp đo lường mức độ phù hợp của các cụm với tập dữ liệu, trong khi những phương pháp khác đo lường mức độ phù hợp của các cụm với sự thật cơ bản, nếu sự thật đó khả dụng. Ngoài ra còn có các biện pháp chấm điểm cụm và do đó có thể so sánh hai tập kết quả cụm trên cùng một tập dữ liệu.

Trong phần còn lại của phần này, chúng ta sẽ thảo luận về từng chủ đề trong ba chủ đề này.

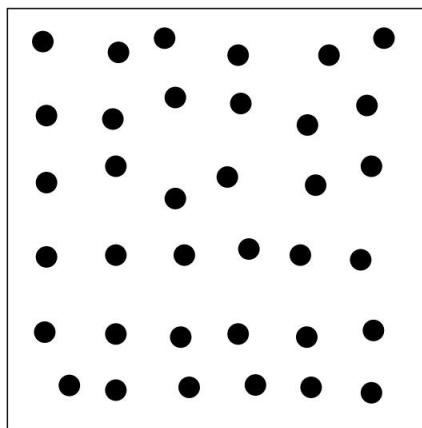
10.6.1 Đánh giá xu hướng cụm

Đánh giá xu hướng phân cụm xác định xem một tập dữ liệu nhất định có cấu trúc không ngẫu nhiên hay không, điều này có thể dẫn đến các cụm có ý nghĩa. Hãy xem xét một tập dữ liệu không có bất kỳ cấu trúc không ngẫu nhiên nào, chẳng hạn như một tập hợp các điểm phân bố đều trong không gian dữ liệu. Mặc dù thuật toán phân cụm có thể trả về các cụm cho dữ liệu, nhưng các cụm đó là ngẫu nhiên và không có ý nghĩa.

Ví dụ 10.9 Phân cụm yêu cầu phân phối dữ liệu không đồng đều. Hình 10.21 cho thấy một tập dữ liệu được phân phối đồng đều trong không gian dữ liệu 2 chiều. Mặc dù thuật toán phân cụm vẫn có thể phân chia các điểm thành các nhóm một cách nhân tạo, nhưng các nhóm này khó có thể có ý nghĩa quan trọng đối với ứng dụng do dữ liệu được phân phối đồng đều. ■

"Làm thế nào chúng ta có thể đánh giá xu hướng phân cụm của một tập dữ liệu?" Theo trực giác, chúng ta có thể thử do xác suất rằng tập dữ liệu được tạo ra bởi một phân phối dữ liệu đồng đều. Điều này có thể đạt được bằng cách sử dụng các bài kiểm tra thống kê về tính ngẫu nhiên không gian. Để minh họa cho ý tưởng này, chúng ta hãy xem xét một thống kê đơn giản nhưng hiệu quả được gọi là Thống kê Hopkins.

Thống kê Hopkins là một thống kê không gian kiểm tra tính ngẫu nhiên không gian của một biến được phân phối trong không gian. Cho một tập dữ liệu, D , được coi là một mẫu



Hình 10.21 Một tập dữ liệu phân bố đồng đều trong không gian dữ liệu.

một biến ngẫu nhiên, σ , chúng ta muốn xác định σ cách xa bao nhiêu so với phân phối đồng đều trong không gian dữ liệu. Chúng ta tính Thống kê Hopkins như sau:

- Lấy mẫu n điểm, p_1, \dots, p_n , đều đặn từ D . Nghĩa là, mỗi điểm trong D có cùng xác suất được đưa vào mẫu này. Đối với mỗi điểm, p_i , chúng ta tìm lân cận gần nhất của p_i ($1 \leq i \leq n$) trong D , và cho x_i là khoảng cách giữa p_i và lân cận gần nhất của nó trong D . Nghĩa là,

$$x_i = \min_{v \in D} \{dist(p_i, v)\}. \quad (10.25)$$

- Lấy mẫu n điểm, q_1, \dots, q_n , đều đặn từ D . Đối với mỗi q_i ($1 \leq i \leq n$), chúng ta tìm hàng xóm gần nhất của q_i trong $D - \{q_i\}$, và cho y_i là khoảng cách giữa q_i và hàng xóm gần nhất của nó trong $D - \{q_i\}$. Nghĩa là,

$$y_i = \min_{v \in D, v \neq q_i} \{dist(q_i, v)\}. \quad (10.26)$$

- Tính Thống kê Hopkins, H , như sau

$$H = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N y_i}. \quad (10.27)$$

"Thống kê Hopkins cho chúng ta biết điều gì về khả năng tập dữ liệu D tuân theo một uniform phân phối trong không gian dữ liệu?" Nếu D được phân phối đều, thì $\sum_{i=1}^N x_i$ và $\sum_{i=1}^N y_i$ sẽ gần nhau, và do đó H sẽ bằng khoảng $0,5$. Tuy nhiên, nếu D là $\sum_{i=1}^N x_i$ trong kỳ vọng, bị lệch rất nhiều, khi đó $\sum_{i=1}^N y_i$ sẽ nhỏ hơn đáng kể và do đó H sẽ gần bằng 0 .

Giả thuyết không của chúng tôi là giả thuyết đồng nhất–rằng D được phân phôi đồng đều và do đó không chứa các cụm có ý nghĩa. Giả thuyết không đồng nhất (tức là D không được phân phôi đồng đều và do đó chứa các cụm) là giả thuyết thay thế.

Chúng ta có thể tiến hành kiểm định Thống kê Hopkins theo cách lặp lại, sử dụng $\theta_0 = 0,5$ làm ngưỡng để bác bỏ giả thuyết thay thế. Nghĩa là, nếu $H > 0,5$, thì không có khả năng D có các cụm có ý nghĩa thống kê.

10.6.2 Xác định số lượng cụm

Việc xác định số lượng cụm "đúng" trong một tập dữ liệu là rất quan trọng, không chỉ vì một số thuật toán phân cụm như k-means yêu cầu tham số như vậy mà còn vì số lượng cụm thích hợp sẽ kiểm soát độ chi tiết thích hợp của phân tích cụm. Có thể coi đây là việc tìm ra sự cân bằng tốt giữa khả năng nén và độ chính xác trong phân tích cụm. Hãy xem xét hai trường hợp cực đoan. Nếu bạn coi toàn bộ tập dữ liệu là một cụm thì sao? Điều này sẽ tối đa hóa khả năng nén dữ liệu, nhưng phân tích cụm như vậy không có giá trị. Mặt khác, việc coi mỗi đối tượng trong tập dữ liệu là một cụm sẽ mang lại độ phân giải phân cụm tốt nhất (tức là chính xác nhất do khoảng cách bằng không giữa một đối tượng và tâm cụm tương ứng). Trong một số phương pháp như k-means, điều này thậm chí còn đạt được chi phí tốt nhất. Tuy nhiên, việc có một đối tượng trên mỗi cụm không cho phép tóm tắt dữ liệu.

Việc xác định số lượng cụm không hề dễ dàng, thường là do số lượng "đúng" không rõ ràng. Việc tìm ra số lượng cụm phù hợp thường phụ thuộc vào hình dạng và quy mô phân phôi trong tập dữ liệu, cũng như độ phân giải cụm mà người dùng yêu cầu. Có nhiều cách có thể để ước tính số lượng cụm. Ở đây, chúng tôi xin giới thiệu tóm tắt một số phương pháp đơn giản nhưng phổ biến và hiệu quả.

Một phương pháp đơn giản là thiết lập số lượng cụm khoảng $\sqrt{2n}$ điểm. Theo kỳ vọng, mỗi cụm có $\sqrt{2n}$ điểm.

— cho một tập dữ liệu n

Phương pháp khuỷu tay dựa trên quan sát rằng việc tăng số lượng cụm có thể giúp giảm tổng phương sai trong cụm của mỗi cụm. Điều này là do việc có nhiều cụm hơn cho phép nắm bắt các nhóm đối tượng dữ liệu nhỏ hơn, giống nhau hơn. Tuy nhiên, hiệu ứng biên của việc giảm tổng phương sai trong cụm có thể giảm nếu quá nhiều cụm được hình thành, vì việc chia một cụm gần kết thành hai cụm chỉ mang lại sự giảm nhỏ. Do đó, một phương pháp tìm kiếm để chọn đúng số lượng cụm là sử dụng điểm ngoặt trong đường cong của tổng phương sai trong cụm đối với số lượng cụm.

Về mặt kỹ thuật, với một số $k > 0$, chúng ta có thể tạo thành k cụm trên tập dữ liệu đang xét bằng cách sử dụng toán phân cụm như k-means và tính tổng các phương sai trong cụm, $\text{var}(k)$. Sau đó, chúng ta có thể vẽ đường cong của var theo k . Điểm ngoặt đầu tiên (hoặc quan trọng nhất) của đường cong cho thấy số "đúng".

Các phương pháp tiên tiến hơn có thể xác định số lượng cụm bằng cách sử dụng tiêu chí thông tin hoặc các phương pháp tiếp cận lý thuyết thông tin. Vui lòng tham khảo các ghi chú thư mục để biết thêm thông tin (Phần 10.9).

Số lượng cụm "đúng" trong một tập dữ liệu cũng có thể được xác định bằng xác thực chéo, một kỹ thuật thường được sử dụng trong phân loại (Chương 8). Đầu tiên, chia tập dữ liệu đã cho, D, thành m phần. Tiếp theo, sử dụng m – 1 phần để xây dựng mô hình cụm và sử dụng phần còn lại để kiểm tra chất lượng cụm. Ví dụ, đối với mỗi điểm trong tập kiểm tra, chúng ta có thể tìm thấy trọng tâm gần nhất. Do đó, chúng ta có thể sử dụng tổng bình phương khoảng cách giữa tất cả các điểm trong tập kiểm tra và các trọng tâm gần nhất để đo mức độ phù hợp của mô hình cụm với tập kiểm tra. Đối với bất kỳ số nguyên k > 0 nào, chúng ta lặp lại quy trình này m lần để suy ra cụm của k cụm bằng cách sử dụng từng phần theo lượt làm tập kiểm tra.

Giá trị trung bình của thước đo chất lượng được coi là thước đo chất lượng tổng thể. Sau đó, chúng ta có thể so sánh thước đo chất lượng tổng thể với các giá trị k khác nhau và tìm ra số cụm phù hợp nhất với dữ liệu.

10.6.3 Đo lường chất lượng cụm

Giả sử bạn đã đánh giá xu hướng phân cụm của một tập dữ liệu nhất định. Bạn cũng có thể đã thử xác định trước số lượng cụm trong tập dữ liệu. Vậy giờ bạn có thể áp dụng một hoặc nhiều phương pháp phân cụm để có được các cụm của tập dữ liệu. "Cụm do một phương pháp tạo ra tốt như thế nào và làm thế nào chúng ta có thể so sánh các cụm do các phương pháp khác nhau tạo ra?"

Chúng ta có một số phương pháp để lựa chọn nhằm đo lường chất lượng của một cụm.

Nhìn chung, các phương pháp này có thể được phân loại thành hai nhóm tùy theo sự thật cơ bản có sẵn hay không. Ở đây, sự thật cơ bản là cụm lý tưởng thường được xây dựng bằng cách sử dụng các chuyên gia con người.

Nếu có sẵn sự thật cơ bản, chúng ta có thể sử dụng nó bằng các phương pháp bên ngoài, so sánh cụm với sự thật nhóm và đo lường. Nếu không có sự thật cơ bản, chúng ta có thể sử dụng các phương pháp bên trong, đánh giá mức độ tốt của cụm bằng cách xem xét các cụm được tách biệt tốt như thế nào. Sự thật cơ bản có thể được coi là giám sát dưới dạng "nhân cụm". Do đó, các phương pháp bên ngoài cũng được gọi là phương pháp có giám sát, trong khi các phương pháp bên trong là phương pháp không có giám sát.

Chúng ta hãy cùng xem xét các phương pháp đơn giản từ mỗi loại.

Phương pháp bên ngoài

Khi có được sự thật cơ bản, chúng ta có thể so sánh nó với một cụm để đánh giá cụm. Do đó, nhiệm vụ cốt lõi trong các phương pháp bên ngoài là gán một điểm, Q(C,Cg), cho một cụm, C, với sự thật cơ bản, Cg. Một phương pháp bên ngoài có hiệu quả hay không phần lớn phụ thuộc vào phép đo, Q, mà nó sử dụng.

Nhìn chung, một biện pháp Q về chất lượng cụm là hiệu quả nếu nó đáp ứng các điều kiện sau: bốn tiêu chí cần thiết:

- Tính đồng nhất của cụm. Điều này đòi hỏi các cụm trong một cụm càng tinh khiết thì cụm càng tốt. Giả sử rằng sự thật cơ bản cho biết các đối tượng trong một tập dữ liệu, D, có thể thuộc về các loại L1,...,Ln. Hãy xem xét cụm, C1, trong đó một cụm C – C1 chứa các đối tượng từ hai loại Li ,Lj (1 ≤ i < j ≤ n). Ngoài ra

Hãy xem xét cụm C2, giống hệt với C1 ngoại trừ việc C2 được chia thành hai cụm chứa các đối tượng trong L1 và Lj tương ứng. Một biện pháp chất lượng cụm, Q, liên quan đến tính đồng nhất của cụm sẽ cho điểm cao hơn cho C2 so với C1, tức là $Q(C2, Cg) > Q(C1, Cg)$.

- Độ hoàn chỉnh của cụm. Đây là đối ứng của tính đồng nhất của cụm. Độ hoàn chỉnh của cụm yêu cầu rằng đối với một cụm, nếu bất kỳ hai đối tượng nào thuộc cùng một loại theo sự thật cơ bản, thì chúng phải được gán vào cùng một cụm. Độ hoàn chỉnh của cụm yêu cầu rằng một cụm phải gán các đối tượng thuộc cùng một loại (theo sự thật cơ bản) vào cùng một cụm. Hãy xem xét cụm C1, bao gồm các cụm C1 và C2, trong đó các thành viên thuộc cùng một loại theo sự thật cơ bản. Giả sử cụm C2 giống hệt với C1 ngoại trừ C1 và C2 được hợp nhất thành một cụm trong C2. Sau đó, một biện pháp chất lượng cụm, Q, liên quan đến độ hoàn chỉnh của cụm sẽ cho điểm cao hơn cho C2, tức là $Q(C2, Cg) > Q(C1, Cg)$.
- Túi giẻ rách. Trong nhiều tình huống thực tế, thường có một danh mục "túi giẻ rách" chứa các đối tượng không thể hợp nhất với các đối tượng khác. Một danh mục như vậy thường được gọi là "hỗn tạp", "khác", v.v. Tiêu chí túi giẻ rách nêu rằng việc đưa một đối tượng không đồng nhất về mặt heterogeneous vào một cụm thuận túy sẽ bị phạt nhiều hơn so với việc đưa nó vào một túi giẻ rách. Hãy xem xét một cụm C1 và một cụm C = C1 sao cho tất cả các đối tượng trong C ngoại trừ một đối tượng, được ký hiệu là o, đều thuộc cùng một danh mục theo sự thật cơ bản. Hãy xem xét một cụm C2 giống hệt với C1 ngoại trừ o được gán cho một cụm C = C trong C2 sao cho C chứa các đối tượng từ nhiều loại khác nhau theo sự thật cơ bản và do đó là nhiều. Nói cách khác, C trong C2 là một túi giẻ rách. Sau đó, một biện pháp chất lượng cụm Q tôn trọng tiêu chí túi giẻ rách sẽ cho điểm cao hơn cho C2, tức là $Q(C2, Cg) > Q(C1, Cg)$.
- Bảo toàn cụm nhỏ. Nếu một phạm trù nhỏ được chia thành nhiều phần nhỏ trong một cụm, những phần nhỏ đó có khả năng trở thành nhiều và do đó không thể phát hiện ra phạm trù nhỏ đó từ cụm. Tiêu chuẩn bảo toàn cụm nhỏ nêu rằng việc chia một phạm trù nhỏ thành nhiều phần có hại hơn việc chia một phạm trù lớn thành nhiều phần. Hãy xem xét một trường hợp cực đoan. Giả sử D là một tập dữ liệu gồm $n + 2$ đối tượng sao cho theo sự thật cơ bản, n đối tượng, được ký hiệu là o1, . . . , on, thuộc về một phạm trù và hai đối tượng còn lại, được ký hiệu là on+1, on+2, thuộc về một phạm trù khác. Giả sử cụm C1 có ba cụm, C1 = {o1, . . . , on}, C2 = {on+1} và C3 = {on+2}. Giả sử cụm C2 cũng có ba cụm, cụ thể là C1 = {o1, . . . , on-1}, C2 = {on}, và C3 = {on+1, on+2}. Nói cách khác, C1 chia tách danh mục nhỏ và C2 chia tách danh mục lớn. Một biện pháp chất lượng cụm Q bảo toàn các cụm nhỏ sẽ cho điểm cao hơn cho C2, tức là, $Q(C2, Cg) > Q(C1, Cg)$.

Nhiều biện pháp chất lượng phân cụm đáp ứng một số trong bốn tiêu chí này. Ở đây, chúng tôi giới thiệu các số liệu độ chính xác và thu hồi BCubed, đáp ứng cả bốn tiêu chí.

BCubed đánh giá độ chính xác và độ thu hồi cho mọi đối tượng trong một cụm trên một tập dữ liệu nhất định theo sự thật cơ bản. Độ chính xác của một đối tượng cho biết có bao nhiêu đối tượng khác trong cùng một cụm thuộc cùng một danh mục với đối tượng đó. Độ thu hồi

của một đối tượng phản ánh có bao nhiêu đối tượng cùng loại được gán vào cùng một cụm.

Về mặt hình thức, hãy để $D = \{o_1, \dots, o_n\}$ là một tập hợp các đối tượng và C là một cụm trên D . Hãy để $L(o_i)$ ($1 \leq i \leq n$) là phạm trù của o_i được đưa ra bởi sự thật cơ bản và $C(o_i)$ là ID cụm của o_i trong C . Khi đó, đối với hai đối tượng, o_i và o_j , ($1 \leq i, j \leq n, i = j$), tính đúng đắn của mối quan hệ giữa o_i và o_j trong cụm C được đưa ra bởi

$$\text{Độ chính xác}(o_i, o_j) = \begin{cases} 1 & \text{nếu } L(o_i) = L(o_j) \quad C(o_i) = C(o_j) \\ 0 & \text{nếu không.} \end{cases} \quad (10.28)$$

Độ chính xác BCubed được định nghĩa là

$$\text{BCubed chính xác} = \frac{\sum_{\substack{o_j : i=j, C(o_i)=C(o_j)}} \text{Độ chính xác}(o_i, o_j)}{\sum_{\substack{o_j | i = j, C(o_i) = C(o_j)}} N}. \quad (10.29)$$

Thu hồi BCubed được định nghĩa là

$$\text{Nhớ lại BCubed} = \frac{\sum_{\substack{o_j : i=j, L(o_i)=L(o_j)}} \text{Độ đúng}(o_i, o_j)}{\sum_{\substack{o_j | i = j, L(o_i) = L(o_j)}} N}. \quad (10.30)$$

Phương pháp nội tại

Khi không có dữ liệu thực tế của một tập dữ liệu, chúng ta phải sử dụng phương pháp nội tại để đánh giá chất lượng phân cụm. Nhìn chung, các phương pháp nội tại đánh giá phân cụm bằng cách kiểm tra mức độ tách biệt của các cụm và mức độ chặt chẽ của các cụm. Nhiều phương pháp nội tại có lợi thế là có số liệu tương tự giữa các đối tượng trong tập dữ liệu.

Hệ số hình bóng là một phép đo như vậy. Đối với một tập dữ liệu, D , gồm n đối tượng, giả sử D được phân chia thành k cụm, C_1, \dots, C_k . Đối với mỗi đối tượng $o \in D$, chúng ta tính $a(o)$ là khoảng cách trung bình giữa o và tất cả các đối tượng khác trong cụm mà o thuộc về.

Tương tự như vậy, $b(o)$ là khoảng cách trung bình nhỏ nhất từ o đến tất cả các cụm mà o không thuộc về. Về mặt hình thức, giả sử C_i ($1 \leq i \leq k$); sau đó

$$a(o) = \frac{\sum_{\substack{o \in C_i, o=o}} \text{phân phối}(o, o)}{|C_i| - 1} \quad (10.31)$$

Và

$$b(o) = \frac{\text{phút}}{\sum_{\substack{o \\ C_j : 1 \leq j \leq k, j=i}}} \cdot \frac{\text{dist}(o, o)}{|C_j|} . \quad (10.32)$$

Hệ số hình bóng của o sau đó được định nghĩa là

$$s(o) = \frac{b(o)}{\max\{a(o), b(o)\}} . \quad (10.33)$$

Giá trị của hệ số hình bóng nằm giữa -1 và 1. Giá trị của $a(o)$ phản ánh tính chật chẽ của cụm mà o thuộc về. Giá trị càng nhỏ, cụm càng chật chẽ. Giá trị của $b(o)$ nằm bắt được mức độ mà o tách biệt khỏi các cụm khác. $b(o)$ càng lớn, o càng tách biệt khỏi các cụm khác. Do đó, khi giá trị hệ số hình bóng của o tiến tới 1, cụm chứa o là chật chẽ và o ở xa các cụm khác, đây là trường hợp được ưu tiên. Tuy nhiên, khi giá trị hệ số hình bóng là âm (tức là $b(o) < a(o)$), điều này có nghĩa là, theo kỳ vọng, o gần với các đối tượng trong cụm khác hơn là các đối tượng trong cùng cụm với o .

Trong nhiều trường hợp, đây là tình huống không tốt và cần phải tránh.

Để đo độ phù hợp của một cụm trong một cụm, chúng ta có thể tính giá trị hệ số hình bóng trung bình của tất cả các đối tượng trong cụm. Để đo chất lượng của một cụm, chúng ta có thể sử dụng giá trị hệ số hình bóng trung bình của tất cả các đối tượng trong tập dữ liệu. Hệ số hình bóng và các phép đo nội tại khác cũng có thể được sử dụng trong phương pháp khuỷu tay để suy ra số cụm trong một tập dữ liệu bằng cách thay thế tổng các phương sai trong cụm.

10.7 Tóm tắt

- Một cụm là một tập hợp các đối tượng dữ liệu giống nhau trong cùng một cụm và khác với các đối tượng trong các cụm khác. Quá trình nhóm một tập hợp các đối tượng vật lý hoặc trừu tượng thành các lớp đối tượng tự được gọi là phân cụm.
- Phân tích cụm có nhiều ứng dụng rộng rãi, bao gồm trí tuệ kinh doanh, nhận dạng mẫu hình ảnh, tìm kiếm trên web, sinh học và bảo mật. Phân tích cụm có thể được sử dụng như một công cụ khai thác dữ liệu độc lập để hiểu sâu hơn về phân phối dữ liệu hoặc như một bước tiền xử lý cho các thuật toán khai thác dữ liệu khác hoạt động trên các cụm đã phát hiện.
- Phân cụm là một lĩnh vực nghiên cứu động trong khai thác dữ liệu. Nó liên quan đến học không giám sát trong học máy.
- Phân cụm là một lĩnh vực đầy thách thức. Các yêu cầu điển hình của nó bao gồm khả năng mở rộng, khả năng xử lý các loại dữ liệu và thuộc tính khác nhau, phát hiện các cụm có hình dạng tùy ý, yêu cầu tối thiểu về kiến thức miền để xác định các tham số đầu vào, khả năng xử lý dữ liệu nhiều, phân cụm gia tăng và

không nhạy cảm với thứ tự đầu vào, khả năng phân cụm dữ liệu có nhiều chiều, phân cụm dựa trên ràng buộc cũng như khả năng diễn giải và khả năng sử dụng.

- Nhiều thuật toán phân cụm đã được phát triển. Chúng có thể được phân loại theo một số khía cạnh trực giao như các khía cạnh liên quan đến tiêu chí phân vùng, phân tách cụm, các biện pháp tương tự được sử dụng và không gian phân cụm. Chương này thảo luận về các phương pháp phân cụm cơ bản chính của các loại sau: phương pháp phân vùng, phương pháp phân cấp, phương pháp dựa trên mật độ và phương pháp dựa trên lưới. Một số thuật toán có thể thuộc về nhiều hơn một loại.
- Phương pháp phân vùng đầu tiên tạo ra một tập hợp ban đầu gồm k phân vùng, trong đó tham số k là số phân vùng cần xây dựng. Sau đó, nó sử dụng kỹ thuật di dời lặp lại nhằm cải thiện phân vùng bằng cách di chuyển các đối tượng từ nhóm này sang nhóm khác. Các phương pháp phân vùng điển hình bao gồm k-means, k-medoids và CLARANS.
- Phương pháp phân cấp tạo ra sự phân tích phân cấp của tập hợp các đối tượng dữ liệu đã cho. Phương pháp này có thể được phân loại là kết tụ (từ dưới lên) hoặc chia tách (từ trên xuống), dựa trên cách phân tích phân cấp được hình thành. Để bù đắp cho tính cứng nhắc của việc hợp nhất hoặc tách, chất lượng của sự kết tụ phân cấp có thể được cải thiện bằng cách phân tích các liên kết đối tượng tại mỗi phân vùng phân cấp (ví dụ, trong Chameleon), hoặc bằng cách đầu tiên thực hiện phân cụm vi mô (tức là nhóm các đối tượng thành "phân cụm vi mô") và sau đó vận hành trên các phân cụm vi mô bằng các kỹ thuật phân cụm khác như di dời lặp lại (như trong BIRCH).
- Phương pháp dựa trên mật độ nhóm các đối tượng dựa trên khái niệm mật độ. Nó phát triển các nhóm theo mật độ của các đối tượng lân cận (ví dụ: trong DBSCAN) hoặc theo hàm mật độ (ví dụ: trong DENCLUE). OPTICS là phương pháp dựa trên mật độ tạo ra thứ tự tăng cường của cấu trúc nhóm dữ liệu.
- Phương pháp dựa trên lưới đầu tiên lượng tử hóa không gian đối tượng thành một số lượng hữu hạn các ô tạo thành cấu trúc lưới, sau đó thực hiện phân cụm trên cấu trúc lưới. STING là một ví dụ điển hình của phương pháp dựa trên lưới dựa trên thông tin thống kê được lưu trữ trong các ô lưới. CLIQUE là thuật toán phân cụm dựa trên lưới và không gian con.
- Đánh giá cụm đánh giá tính khả thi của phân tích cụm trên một tập dữ liệu và chất lượng của kết quả do phương pháp cụm tạo ra. Các nhiệm vụ bao gồm đánh giá xu hướng cụm, xác định số lượng cụm và đo lường chất lượng cụm.

10.8 Bài tập

10.1 Hãy mô tả ngắn gọn và đưa ra ví dụ về từng phương pháp phân cụm sau: phương pháp phân vùng , phương pháp phân cấp , phương pháp dựa trên mật độ và phương pháp dựa trên lưới .

492 Chương 10 Phân tích cụm: Các khái niệm và phương pháp cơ bản

10.2 Giả sử rằng nhiệm vụ khai thác dữ liệu là nhóm các điểm (với (x, y) biểu diễn vị trí) thành ba nhóm, trong đó các điểm là

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

Hàm khoảng cách là khoảng cách Euclidean. Giả sử ban đầu chúng ta gán A1, B1 và C1 làm tâm của mỗi cụm tương ứng. Sử dụng thuật toán k-means để chỉ hiển thị

(a) Ba cụm trung tâm sau vòng thực hiện đầu tiên. (b) Ba cụm cuối cùng.

10.3 Sử dụng một ví dụ để chỉ ra lý do tại sao thuật toán k-means có thể không tìm thấy tối ưu toàn cục, nghĩa là tối ưu hóa sự thay đổi trong cụm.

10.4 Đối với thuật toán k-means, điều thú vị cần lưu ý là bằng cách chọn các trung tâm cụm ban đầu một cách cẩn thận, chúng ta không chỉ có thể tăng tốc độ hội tụ của thuật toán mà còn đảm bảo chất lượng của cụm cuối cùng. Thuật toán k-means++ là một biến thể của k-means, thuật toán này chọn các trung tâm ban đầu như sau. Đầu tiên, thuật toán này chọn một trung tâm một cách đồng đều ngẫu nhiên từ các đối tượng trong tập dữ liệu. Theo chu kỳ, đối với mỗi đối tượng p khác với trung tâm đã chọn, thuật toán sẽ chọn một đối tượng làm trung tâm mới. Đối tượng này được chọn ngẫu nhiên với xác suất tỷ lệ thuận với $dist(p)$ trong đó $dist(p)$ là khoảng cách từ p đến trung tâm gần nhất đã được chọn. Quá trình lặp lại tiếp tục cho đến khi chọn được k trung tâm.

Giải thích tại sao phương pháp này không chỉ tăng tốc độ hội tụ của k-means thuật toán, mà còn đảm bảo chất lượng của kết quả phân cụm cuối cùng.

10.5 Cung cấp mã giả của bước gán lại đối tượng của thuật toán PAM.

10.6 Cả thuật toán k-means và k-medoids đều có thể thực hiện phân cụm hiệu quả.

(a) Minh họa điểm mạnh và điểm yếu của k-means khi so sánh với k-medoids. (b) Minh họa điểm mạnh và điểm yếu của các lược đồ này khi so sánh với lược đồ phân cụ theo thứ bậc (ví dụ: AGNES).

10.7 Chứng minh rằng trong DBSCAN, mối quan hệ mật độ-liên thông là một quan hệ tương đương.

10.8 Chứng minh rằng trong DBSCAN, đối với giá trị MinPts cố định và hai ngưỡng lân cận, $1 < 2$, một cụm C đối với 1 và MinPts phải là một tập hợp con của một cụm C có và MinPts. đối với 2

10.9 Cung cấp mã giả của thuật toán OPTICS.

10.10 Tại sao BIRCH gặp khó khăn trong việc tìm các cụm có hình dạng tùy ý nhưng OPTICS thì không? Đề xuất sửa đổi BIRCH để giúp nó tìm các cụm có hình dạng tùy ý.

10.11 Cung cấp mã giả của bước trong CLIQUE tìm các ô dày đặc trong tất cả các khung gian con.

- 10.12 Các điều kiện hiện tại mà phân cụm dựa trên mật độ phù hợp hơn phân cụm dựa trên phân vùng và phân cụm phân cấp. Đưa ra các ví dụ ứng dụng để hỗ trợ cho lập luận của bạn.
- 10.13 Đưa ra ví dụ về cách tích hợp các phương pháp phân cụm cụ thể, ví dụ, khi một thuật toán phân cụm được sử dụng làm bước tiền xử lý cho một thuật toán khác. Ngoài ra, hãy đưa ra lý do tại sao việc tích hợp hai phương pháp đôi khi có thể dẫn đến cải thiện chất lượng và hiệu quả phân cụm.
- 10.14 Phân cụm được công nhận là một nhiệm vụ khai thác dữ liệu quan trọng với các ứng dụng rộng rãi. Đưa ra một ví dụ ứng dụng cho mỗi trường hợp sau:
- (a) Một ứng dụng sử dụng cụm như một chức năng khai thác dữ liệu chính. (b) Một ứng dụng sử dụng cụm như một công cụ xử lý trước để chuẩn bị dữ liệu cho các tác vụ khai thác dữ liệu khác.
- 10.15 Khái dữ liệu và cơ sở dữ liệu đa chiều chứa dữ liệu danh nghĩa, thứ tự và số ở dạng phân cấp hoặc tổng hợp. Dựa trên những gì bạn đã học về các phương pháp phân cụm, hãy thiết kế một phương pháp phân cụm tìm các cụm trong khối dữ liệu lớn một cách hiệu quả.
- 10.16 Mô tả từng thuật toán phân cụm sau đây theo các tiêu chí sau: (1) hình dạng của các cụm có thể xác định được; (2) các tham số đầu vào phải được chỉ định; và (3) các hạn chế. (a) k-means (b) k-medoids (c) CLARA (d) BIRCH
 (e) CHAMELEON
 (f) DBSCAN
- 10.17 Mắt người có thể đánh giá nhanh và hiệu quả chất lượng của các phương pháp phân cụm dữ liệu 2 chiều. Bạn có thể thiết kế một phương pháp trực quan hóa dữ liệu có thể giúp con người trực quan hóa các cụm dữ liệu và đánh giá chất lượng phân cụm dữ liệu 3 chiều không? Còn dữ liệu có chiều cao hơn thì sao?
- 10.18 Giả sử bạn phân bổ một số máy ATM trong một khu vực nhất định để đáp ứng một số ràng buộc. Các hộ gia đình hoặc nơi làm việc có thể được nhóm lại sao cho thông thường một máy ATM được chỉ định cho mỗi nhóm. Tuy nhiên, việc nhóm lại có thể bị hạn chế bởi hai yếu tố: (1) các vật cản (tức là có cầu, sông và đường cao tốc có thể ảnh hưởng đến khả năng tiếp cận ATM) và (2) các ràng buộc bổ sung do người dùng chỉ định như mỗi máy ATM phải phục vụ ít nhất 10.000 hộ gia đình. Làm thế nào để một thuật toán nhóm lại như k-means có thể được sửa đổi để nhóm lại chất lượng theo cả hai ràng buộc?
- 10.19 Đối với cụm dựa trên ràng buộc, ngoài việc có số lượng khách hàng tối thiểu trong mỗi cụm (để phân bổ ATM) làm ràng buộc, có thể có nhiều loại khác

ràng buộc. Ví dụ, một ràng buộc có thể ở dạng số lượng khách hàng tối đa trên mỗi cụm, thu nhập trung bình của khách hàng trên mỗi cụm, khoảng cách tối đa giữa mỗi hai cụm, v.v. Phân loại các loại ràng buộc có thể áp dụng cho các cụm được tạo ra và thảo luận về cách thực hiện phân cụm hiệu quả theo các loại ràng buộc đó.

- 10.20 Thiết kế phương pháp phân cụ đảm bảo quyền riêng tư để chủ sở hữu dữ liệu có thể yêu cầu bên thứ ba khai thác dữ liệu nhằm phân cụ chất lượng mà không phải lo lắng về khả năng tiết lộ không phù hợp một số thông tin riêng tư hoặc nhạy cảm được lưu trữ trong dữ liệu.
- 10.21 Chứng minh rằng các số liệu BCubed đáp ứng bốn yêu cầu thiết yếu cho phân cụm bên ngoài phương pháp đánh giá.

10.9 Ghi chú tài liệu tham khảo

Phân cụm đã được nghiên cứu rộng rãi trong hơn 40 năm và trên nhiều lĩnh vực do ứng dụng rộng rãi của nó. Hầu hết các sách về phân loại mẫu và học máy đều có các chương về phân tích cụm hoặc học không giám sát. Một số sách giáo khoa dành riêng cho các phương pháp phân tích cụm, bao gồm Hartigan [Har75]; Jain và Dubes [JD88]; Kaufman và Rousseeuw [KR90]; và Arabie, Hubert và De Soete [AHS96]. Ngoài ra còn có nhiều bài báo khảo sát về các khía cạnh khác nhau của phương pháp phân cụm.

Những người gần đây bao gồm Jain, Murty và Flynn [JMF99]; Parsons, Haque và Liu [PHL04]; và Jain [Jai10].

Đối với các phương pháp phân vùng, thuật toán k-means đầu tiên được Lloyd [Llo57] giới thiệu, và sau đó là MacQueen [Mac67]. Arthur và Vassilvitskii [AV07] đã trình bày thuật toán k-means++. Một thuật toán lọc, sử dụng chỉ số dữ liệu phân cấp không gian để tăng tốc tính toán các giá trị trung bình cụm, được đưa ra trong Kanungo, Mount, Netanyahu, et al. [KMN+02].

Thuật toán k-medoids của PAM và CLARA được đề xuất bởi Kaufman và Rousseeuw [KR90]. Thuật toán k-modes (để phân cụm dữ liệu danh nghĩa) và k-prototypes (để phân cụm dữ liệu lai) được đề xuất bởi Huang [Hua98]. Thuật toán phân cụm k-modes cũng được đề xuất độc lập bởi Chaturvedi, Green và Carroll [CGC94, CGC01]. Thuật toán CLARANS được đề xuất bởi Ng và Han [NH94].

Ester, Kriegel và Xu [EKX95] đã đề xuất các kỹ thuật để cải thiện hiệu suất của CLARANS hơn nữa bằng cách sử dụng các phương pháp truy cập không gian hiệu quả như R-tree và các kỹ thuật tập trung. Một thuật toán phân cụ có thể mở rộng dựa trên k-means đã được Bradley, Fayyad và Reina [BFR98] đề xuất.

Một khảo sát ban đầu về các thuật toán phân cụ phân cấp kết tụ đã được Day và Edelsbrunner [DE84] thực hiện. Phân cụ phân cấp kết tụ, chẳng hạn như AGNES, và phân cụ phân cấp chia tách, chẳng hạn như DIANA, đã được Kaufman và Rousseeuw [KR90] giới thiệu. Một hướng thú vị để cải thiện chất lượng phân cụ của các phương pháp phân cụ phân cấp là tích hợp phân cụ phân cấp với dịch chuyển lặp lại dựa trên khoảng cách hoặc các phương pháp phân cụ phi phân cấp khác. Ví dụ, BIRCH, của Zhang, Ramakrishnan và Livny [ZRL96], đầu tiên thực hiện phân cụ phân cấp với

một cây CF trước khi áp dụng các kỹ thuật khác. Phân cụm phân cấp cũng có thể được thực hiện bằng phân tích liên kết tinh vi, chuyển đổi hoặc phân tích lân cận gần nhất, chẳng hạn như CURE của Guha, Rastogi và Shim [GRS98]; ROCK (để nhóm các danh nghĩa thuộc tính) của Guha, Rastogi và Shim [GRS99]; và Chameleon của Karypis, Han và Kumar [KHK99].

Một khuôn khổ phân cụm xác suất theo thuật toán liên kết thông thường và sử dụng các mô hình xác suất để xác định độ tương đồng của cụm đã được phát triển bởi Friedman [Fri03] và Heller và Ghahramani [HG05].

Đối với các phương pháp phân cụm dựa trên mật độ, DBSCAN được đề xuất bởi Ester, Kriegel, Sander và Xu [EKSX96]. Ankerst, Breunig, Kriegel và Sander [ABKS99] đã phát triển OPTICS, một phương pháp sắp xếp cụm giúp tạo điều kiện cho việc phân cụm dựa trên mật độ mà không cần lo lắng về thông số kỹ thuật. Thuật toán DENCLUE, dựa trên một tập hợp của các hàm phân phối mật độ, được đề xuất bởi Hinneburg và Keim [HK98]. Hinneburg và Gabriel [HG07] đã phát triển DENCLUE 2.0, bao gồm một quy trình leo dốc cho hạt nhân Gaussian tự động điều chỉnh kích thước bước.

STING, một phương pháp tiếp cận đa độ phân giải dựa trên lưới thu thập thông tin thống kê trong các ô lưới, được đề xuất bởi Wang, Yang và Muntz [WYM97]. WaveCluster, được phát triển bởi Sheikholeslami, Chatterjee và Zhang [SCZ98], là một cụm phân giải đa phương pháp biến đổi không gian đặc trưng ban đầu bằng phép biến đổi wavelet.

Các phương pháp có thể mở rộng để phân cụm dữ liệu danh nghĩa đã được nghiên cứu bởi Gibson, Kleinberg, và Raghavan [GKR98]; Guha, Rastogi, và Shim [GRS99]; và Ganti, Gehrke, và Ramakrishnan [GGR99]. Ngoài ra còn có nhiều mô hình cụm khác. Ví dụ, các phương pháp cụm mở được thảo luận trong Kaufman và Rousseeuw [KR90], Bezdek [Bez81], và Bezdek và Pal [BP92].

Đối với cụm chiều cao, một thuật toán cụm không gian con tăng trưởng chiều dựa trên Apriori có tên là CLIQUE đã được đề xuất bởi Agrawal, Gehrke, Gunopulos và Raghavan [AGGR98]. Nó tích hợp các phương pháp phân cụm dựa trên mật độ và dựa trên lưới.

Các nghiên cứu gần đây đã tiến hành phân cụm dữ liệu luồng Babcock, Badu, Datar, et al. [BBD+02]. Một thuật toán phân cụm luồng dữ liệu dựa trên k-trung vị đã được đề xuất bởi Guha, Mishra, Motwani, và O'Callaghan [GMM00] và bởi O'Callaghan et al. [OMM+02].

Một phương pháp để nhóm các luồng dữ liệu đang phát triển đã được đề xuất bởi Aggarwal, Han, Wang, và Yu [AHWY03]. Một khuôn khổ cho việc phân cụm dữ liệu có chiều cao được dự kiến luồng được đề xuất bởi Aggarwal, Han, Wang và Yu [AHWY04a].

Đánh giá cụm được thảo luận trong một số chuyên khảo và bài báo khảo sát như Jain và Dubes [JD88] và Halkidi, Batistakis và Vazirgiannis [HBV01]. Các phương pháp bên ngoài để đánh giá chất lượng cụm được khám phá rộng rãi. Một số nghiên cứu gần đây bao gồm Meila [Mei03, Mei05] và Amigo, Gonzalo, Artiles và Verdejo [AGAV09].

Bốn tiêu chí thiết yếu được giới thiệu trong chương này được xây dựng trong Amigo, Gonzalo, Artiles và Verdejo [AGAV09], trong khi một số tiêu chí riêng lẻ cũng được đề cập trước đó, ví dụ, trong Meila [Mei03] và Rosenberg và Hirschberg [RH07]. Baggà và Baldwin [BB98] đã giới thiệu các số liệu BCubed. Hệ số hình bóng được mô tả trong Kaufman và Rousseeuw [KR90].

Trang này có ý đẽ trống