Mục lục

Chương 1

Giới thiệu

Cuốn sách này là một lời giới thiệu về lĩnh vực *khai phá dữ liệu* trẻ trung và phát triển nhanh (còn được biết đến với tên gọi *Khám phá tri thức từ dữ liệu*, hay viết tắt là *KDD*). Cuốn sách tập trung vào các khái niệm và kỹ thuật cơ bản của khai phá dữ liệu nhằm phát hiện các mẫu thú vị từ dữ liệu trong nhiều ứng dụng khác nhau. Đặc biệt, chúng tôi nhấn mạnh các kỹ thuật nổi bật để phát triển các công cụ khai phá dữ liệu hiệu quả, hiệu suất cao và có khả năng mở rộng.

Chương này được sắp xếp như sau. Ở Mục 1.1, chúng ta tìm hiểu khai phá dữ liệu là gì và tại sao nó lại có nhu cầu cao. Mục 1.2 liên kết khai phá dữ liệu với quy trình khám phá tri thức tổng thể. Tiếp theo, chúng ta sẽ tìm hiểu về khai phá dữ liệu từ nhiều khía cạnh khác nhau, chẳng hạn như các loại dữ liệu có thể được khai phá (Mục 1.3), các loại tri thức cần được phát hiện (Mục 1.4), mối quan hệ giữa

khai phá dữ liệu với các ngành khoa học khác (Mục 1.5) và các ứng dụng của khai phá dữ liệu (Mục 1.6). Cuối cùng, chúng ta thảo luận về tác động của khai phá dữ liệu đối với xã hội (Mục 1.7).

1.1 Khai phá dữ liệu là gì?

Nhu cầu là nguồn gốc của phát minh
– Plato

Chúng ta đang sống trong một thế giới mà dữ liệu được tạo ra liên tục và với tốc độ nhanh chóng.

"*Chúng ta đang sống trong kỷ nguyên thông tin*" là một câu nói phổ biến; tuy nhiên, *thực tế chúng ta đang sống trong kỷ nguyên dữ liệu*. Mỗi ngày, hàng terabyte hoặc petabyte dữ liệu đổ vào các mạng máy tính, World Wide Web (WWW) và nhiều loại thiết bị khác nhau từ các lĩnh vực kinh doanh, báo chí, xã hội, khoa học, kỹ thuật, y học và hầu hết mọi khía cạnh của cuộc sống hàng ngày. Sự gia tăng bùng nổ về khối lượng dữ liệu sẵn có này là kết quả của quá trình tin học hóa xã hội cùng với sự phát triển nhanh chóng của các công cụ tính toán mạnh mẽ, cảm biến, thu thập dữ liệu, lưu trữ và xuất bản thông tin.

Các doanh nghiệp trên toàn cầu tạo ra những tập dữ liệu khổng lồ, bao gồm giao dịch bán hàng, hồ sơ giao dịch chứng khoán, mô tả sản phẩm, chương trình khuyến mãi, hồ sơ công ty và hiệu suất hoạt động, cũng như phản hồi từ khách hàng. Các hoạt động khoa học và kỹ thuật liên tục tạo ra dữ liệu với quy mô hàng petabyte, từ cảm biến từ xa, đo lường quy trình, thí nghiệm khoa học, đánh giá hiệu suất hệ thống, quan sát kỹ thuật đến giám sát môi trường. Lĩnh vực nghiên cứu y sinh và công nghiệp y tế tạo ra một lượng dữ liệu khổng lồ từ các máy giải trình tự gen, báo cáo thí nghiệm và nghiên cứu y sinh, hồ sơ bệnh án, giám sát bệnh nhân và hình ảnh y khoa. Hàng tỷ lượt tìm kiếm trên Web được các công cụ tìm kiếm xử lý mỗi ngày với hàng chục petabyte dữ liệu. Các công cụ truyền thông xã hội ngày càng trở nên phổ biến, tạo ra vô số văn bản, hình ảnh và video, đồng thời hình thành nhiều cộng đồng Web và mạng xã hội khác nhau. Danh sách các nguồn dữ liệu khổng lồ này gần như là vô tận.

Sự bùng nổ về dữ liệu khổng lồ và phổ biến này đã biến thời đại của chúng ta thành *kỷ nguyên dữ liệu* thực sự. Các công cụ mạnh mẽ và đa năng là điều cấp thiết để tự động phát hiện những thông tin quý giá từ khối lượng dữ liệu khổng lồ

và biến chúng thành tri thức có tổ chức. Chính nhu cầu đó đã dẫn đến sự ra đời của khai phá dữ liệu.

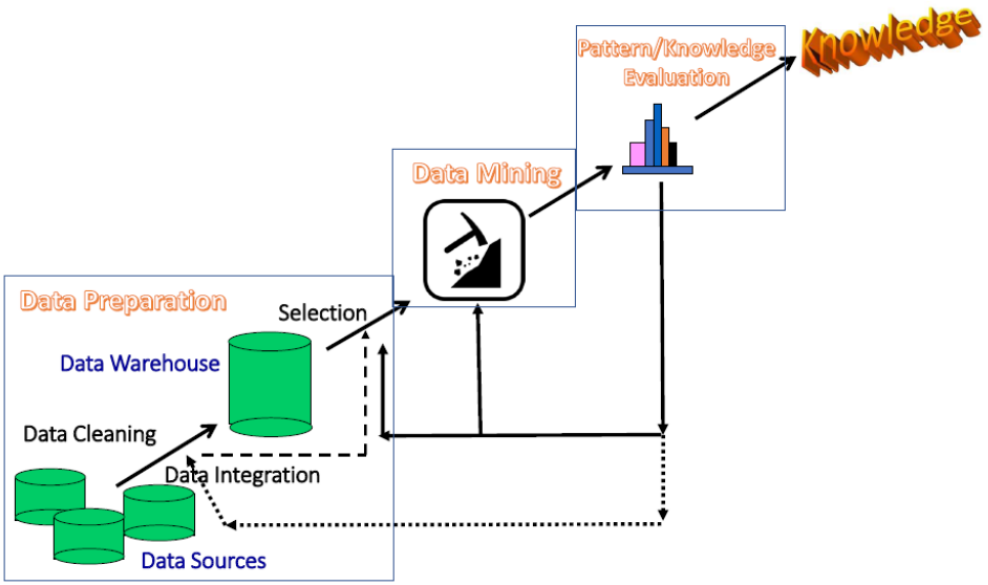
Về cơ bản, ***khai phá dữ liệu*** *là quá trình phát hiện các mẫu, mô hình và những loại tri thức khác thú vị trong các tập dữ liệu lớn*. Thuật ngữ "*khai phá dữ liệu*" ra đời vào những năm 1990, gợi lên hình ảnh tìm kiếm những *mẩu vàng quý giá* từ dữ liệu. Tuy nhiên, khi nói đến việc khai thác vàng từ đá hay cát, chúng ta thường sử dụng thuật ngữ "*đào vàng*" hay "*khai thác vàng*" thay vì "khai thác đá" hay "khai thác cát". Tương tự, khai phá dữ liệu lẽ ra nên được gọi là "khai thác tri thức từ dữ liệu", tuy nhiên tên gọi này lại hơi dài dòng. Dù sao, thuật ngữ ngắn gọn "*khai phá dữ liệu*" cũng đủ phản ánh quá trình tìm kiếm một tập hợp nhỏ những mẩu vàng quý giá từ một khối lượng lớn nguyên liệu thô. Do đó, cái tên có vẻ "sai lệch" nhưng mang cả "dữ liệu" và "khai thác" đã trở thành lựa chọn phổ biến. Ngoài ra, còn có nhiều thuật ngữ khác có ý nghĩa tương tự như khai phá dữ liệu — ví dụ như *khai thác tri thức từ dữ liệu, KDD* (viết tắt của *Knowledge Discovery from Data*), *phát hiện mẫu*, *khám phá mẫu*, *trích xuất tri thức*, *khảo cổ dữ liệu*, *phân tích dữ liệu* và *thu hoạch thông tin*.

Khai phá dữ liệu là một lĩnh vực trẻ trung, năng động và đầy hứa hẹn. Nó đã và sẽ tiếp tục tạo ra những bước tiến vượt bậc trên hành trình chuyển mình từ kỷ nguyên dữ liệu sang kỷ nguyên thông tin sắp tới.

Ví dụ 1 (Khai phá dữ liệu biến một tập hợp lớn dữ liệu thành tri thức)**.** Một công cụ tìm kiếm (ví dụ, Google) nhận được hàng tỷ truy vấn mỗi ngày. Những tri thức mới lạ và hữu ích nào mà một công cụ tìm kiếm có thể học được từ một tập hợp khổng lồ các truy vấn được thu thập từ người dùng theo thời gian? Thật thú vị, một số mẫu được phát hiện trong các truy vấn tìm kiếm của người dùng có thể tiết lộ những tri thức vô giá mà không thể thu được chỉ bằng cách đọc từng mục dữ liệu riêng lẻ. Ví dụ, *Flu Trends* của Google sử dụng các từ khóa cụ thể như các chỉ số cho diễn biến của bệnh cúm. Họ phát hiện ra mối quan hệ chặt chẽ giữa số lượng người tìm kiếm thông tin liên quan đến cúm và số lượng người thực sự có triệu chứng cúm. Một mẫu nhất định xuất hiện khi tất cả các truy vấn tìm kiếm liên quan đến cúm được tổng hợp lại. Dựa vào dữ liệu tìm kiếm được tổng hợp của Google, *Flu Trends* có thể ước tính diễn biến của bệnh cúm nhanh hơn đến hai tuần so với các hệ thống truyền thống.* Ví dụ này cho thấy khai phá dữ liệu có thể biến một tập hợp lớn dữ liệu thành tri thức hữu ích, góp phần giải quyết những thách thức toàn cầu hiện nay.

1.2 Khai phá dữ liệu: bước thiết yếu trong khám phá tri thức

Nhiều người xem khai phá dữ liệu như một từ đồng nghĩa với thuật ngữ phổ biến khác, đó là **khám phá tri thức từ dữ liệu** (hay **KDD**), trong khi một số khác lại coi khai phá dữ liệu chỉ là một bước thiết yếu trong quy trình khám phá tri thức tổng thể. Quy trình khám phá tri thức tổng thể được thể hiện trong Hình 1.1 như một chuỗi lặp đi lặp lại các bước sau:



Hình 1.1: Khai phá dữ liệu: bước thiết yếu trong quy trình khám phá tri thức.

1) a) **Làm sạch dữ liệu** (để loại bỏ nhiễu và dữ liệu không nhất quán)

   b) **Tích hợp dữ liệu** (nơi có thể kết hợp nhiều nguồn dữ liệu)*

   c) **Biến đổi dữ liệu** (nơi dữ liệu được chuyển đổi và hợp nhất thành các dạng phù hợp cho việc khai phá thông qua các thao tác tóm tắt hoặc tổng hợp)†

*Điều này đã được báo cáo trong *"Detecting influenza epidemics using search engine query data"* (2009). Báo cáo của *Flu Trends* dừng vào năm 2015.

*Một xu hướng phổ biến trong ngành thông tin là thực hiện làm sạch và tích hợp dữ liệu như một bước tiền xử lý, trong đó dữ liệu thu được sau đó được lưu trữ vào một kho dữ liệu.

†Việc biến đổi và hợp nhất dữ liệu thường được thực hiện trước quá trình chọn lọc dữ liệu, đặc biệt trong trường hợp của kho dữ liệu. Việc *giảm dữ liệu* cũng có thể được thực hiện để có được một biểu diễn nhỏ gọn hơn của dữ liệu gốc mà không làm mất đi tính toàn vẹn của nó.

d) **Chọn lọc dữ liệu** (nơi lấy ra các dữ liệu liên quan đến nhiệm vụ phân tích từ cơ sở dữ liệu)

2) **Khai phá dữ liệu** (một quá trình thiết yếu, nơi các phương pháp thông minh được áp dụng để trích xuất các mẫu hoặc xây dựng các mô hình)

3) **Đánh giá mẫu / mô hình** (để xác định các mẫu hoặc mô hình thực sự thú vị đại diện cho tri thức dựa trên các chỉ số *đo lường độ thú vị* — xem Mục 1.4.7)

4) **Trình bày tri thức** (nơi các kỹ thuật trực quan hóa và biểu diễn tri thức được sử dụng để trình bày tri thức đã khai phá cho người dùng)

Các bước 1(a) đến 1(d) là các hình thức tiền xử lý dữ liệu khác nhau, trong đó dữ liệu được chuẩn bị cho việc khai phá. Bước khai phá dữ liệu có thể tương tác với người dùng hoặc với cơ sở tri thức. Các mẫu thú vị sau đó được trình bày cho người dùng và có thể được lưu trữ như tri thức mới trong cơ sở tri thức.

Quan điểm trên cho thấy khai phá dữ liệu chỉ là một bước trong quy trình khám phá tri thức, mặc dù đó là bước thiết yếu vì nó khai phá những mẫu hoặc mô hình ẩn để đánh giá. Tuy nhiên, trong ngành công nghiệp, truyền thông và cả trong giới nghiên cứu, thuật ngữ *khai phá dữ liệu* thường được dùng để chỉ toàn bộ quy trình khám phá tri thức (có lẽ vì thuật ngữ này ngắn gọn hơn *khám phá tri thức từ dữ liệu*). Do đó, chúng ta sẽ áp dụng một quan điểm rộng về chức năng của khai phá dữ liệu: *Khai phá dữ liệu là quá trình phát hiện các mẫu và tri thức thú vị từ khối lượng dữ liệu lớn*. Các nguồn dữ liệu có thể bao gồm cơ sở dữ liệu, kho dữ liệu, Web, các kho lưu trữ thông tin khác, hoặc dữ liệu được truyền động vào hệ thống theo thời gian thực.

1.3 Tính đa dạng của kiểu dữ liệu trong khai phá dữ liệu

Là một công nghệ chung, khai phá dữ liệu có thể được áp dụng cho bất kỳ loại dữ liệu nào miễn là dữ liệu đó có ý nghĩa với ứng dụng mục tiêu. Tuy nhiên, các loại dữ liệu khác nhau có thể đòi hỏi các phương pháp khai phá dữ liệu khá khác biệt, từ đơn giản đến phức tạp, tạo nên một lĩnh vực phong phú và đa dạng.

1.3.1 Dữ liệu có cấu trúc và dữ liệu không có cấu trúc

Dựa trên việc dữ liệu có cấu trúc rõ ràng hay không, ta có thể phân loại dữ liệu thành *dữ liệu có cấu trúc* và *dữ liệu không có cấu trúc*.

1.3.1.1 Dữ liệu có cấu trúc

Dữ liệu được lưu trữ *trong cơ sở dữ liệu quan hệ*, *dữ liệu khối*, *ma trận dữ liệu* và nhiều *kho dữ liệu* thường có cấu trúc đồng nhất, dạng bản ghi hoặc bảng, được định nghĩa bởi từ điển dữ liệu của chúng với một tập các thuộc tính (hoặc trường, cột) cố định, mỗi thuộc tính có miền giá trị và ý nghĩa ngữ nghĩa xác định. Những tập dữ liệu này là ví dụ điển hình của dữ liệu có cấu trúc cao. Trong nhiều ứng dụng thực tế, yêu cầu cấu trúc chặt chẽ này có thể được nới lỏng theo nhiều cách để phù hợp với tính *bán cấu trúc* của dữ liệu, chẳng hạn như cho phép một đối tượng dữ liệu chứa một giá trị cố định, một tập nhỏ các giá trị có kiểu khác nhau, hoặc các cấu trúc lồng nhau, hoặc cho phép cấu trúc của các đối tượng hoặc đối tượng con được định nghĩa một cách linh hoạt và động (ví dụ, cấu trúc XML).

Có nhiều tập dữ liệu không được cấu trúc chặt chẽ như bảng quan hệ hay ma trận dữ liệu. Tuy nhiên, chúng vẫn có một số cấu trúc với ý nghĩa ngữ nghĩa rõ ràng. Ví dụ, một *tập dữ liệu giao dịch* có thể chứa một tập lớn các giao dịch, mỗi giao dịch bao gồm một tập các mặt hàng. Một *tập dữ liệu chuỗi* có thể chứa một tập lớn các chuỗi, mỗi chuỗi bao gồm một tập các phần tử được sắp xếp theo thứ tự, mà mỗi phần tử lại có thể chứa một tập các mục. Nhiều tập dữ liệu ứng dụng, chẳng hạn như dữ liệu giao dịch mua sắm, dữ liệu chuỗi thời gian, dữ liệu gene hoặc protein, hoặc dữ liệu Weblog, thuộc nhóm này.

Một loại dữ liệu bán cấu trúc phức tạp hơn là *dữ liệu đồ thị hoặc mạng*, nơi một tập các nút được kết nối bởi một tập các cạnh (còn gọi là liên kết); và mỗi nút hoặc liên kết có thể có mô tả ngữ nghĩa hoặc cấu trúc con riêng.

Mỗi loại tập dữ liệu có cấu trúc và bán cấu trúc này có thể chứa những mẫu hoặc tri thức đặc thù cần được khai phá, và nhiều phương pháp khai phá dữ liệu chuyên dụng như khai phá mẫu tuần tự, khai phá mẫu đồ thị và khai phá mạng thông tin đã được phát triển để phân tích các tập dữ liệu này.

1.3.1.2 Dữ liệu không có cấu trúc

Bên cạnh dữ liệu có cấu trúc và bán cấu trúc, còn có một lượng lớn dữ liệu không có cấu trúc, chẳng hạn như dữ liệu văn bản và dữ liệu đa phương tiện (ví dụ: âm thanh, hình ảnh, video). Mặc dù một số nghiên cứu xem chúng như các luồng byte một chiều hoặc nhiều chiều, nhưng chúng chứa đựng rất nhiều ý nghĩa ngữ nghĩa thú vị. Các phương pháp chuyên biệt theo từng lĩnh vực đã được phát triển để phân tích những dữ liệu này trong các lĩnh vực như hiểu ngôn ngữ tự nhiên, khai phá văn bản, thị giác máy tính và nhận dạng mẫu. Hơn nữa, những tiến bộ gần đây

trong học sâu đã đạt được tiến bộ vượt bậc trong xử lý dữ liệu văn bản, hình ảnh và video. Tuy nhiên, việc khai phá các cấu trúc ẩn từ dữ liệu không có cấu trúc có thể giúp hiểu rõ hơn và sử dụng hiệu quả những dữ liệu đó.

1.3.1.3 Dữ liệu hỗn hợp trong thực tế

Trong thế giới thực, dữ liệu thường là sự pha trộn của dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc. Ví dụ, một trang web mua sắm trực tuyến có thể chứa thông tin cho một tập hợp lớn sản phẩm, vốn chủ yếu là dữ liệu có cấu trúc được lưu trữ trong cơ sở dữ liệu quan hệ với một tập các trường cố định như tên sản phẩm, giá cả, thông số kỹ thuật, v.v. Tuy nhiên, một số trường có thể thực chất là dữ liệu văn bản, hình ảnh và video, chẳng hạn như phần giới thiệu sản phẩm, nhận xét của chuyên gia hoặc người dùng, hình ảnh sản phẩm và video quảng cáo. Các phương pháp khai phá dữ liệu thường được phát triển cho việc khai phá một loại dữ liệu cụ thể, và kết quả của chúng có thể được tích hợp và phối hợp để phục vụ mục tiêu chung.

1.3.2 Dữ liệu liên kết với các ứng dụng khác nhau

Các ứng dụng khác nhau có thể tạo ra hoặc cần xử lý các tập dữ liệu rất khác nhau và đòi hỏi các phương pháp phân tích dữ liệu cũng khác biệt. Vì vậy, khi phân loại các tập dữ liệu cho khai phá dữ liệu, chúng ta nên cân nhắc đến các ứng dụng cụ thể.

Lấy dữ liệu chuỗi làm ví dụ. Các *chuỗi sinh học* như chuỗi DNA hoặc protein có thể mang ý nghĩa ngữ nghĩa rất khác so với các *chuỗi giao dịch mua sắm* hoặc *luồng nhấp chuột trên Web*, đòi hỏi các phương pháp khai phá chuỗi khác nhau. Một loại dữ liệu chuỗi đặc biệt là dữ liệu *chuỗi thời gian*, trong đó một chuỗi thời gian có thể chứa một tập hợp các giá trị số được sắp thứ tự với khoảng thời gian bằng nhau, điều này cũng khá khác so với chuỗi giao dịch mua sắm, vốn không nhất thiết có khoảng cách thời gian cố định (một khách hàng có thể mua sắm bất cứ lúc nào người đó thích).

Dữ liệu trong một số ứng dụng có thể đi kèm với thông tin không gian, thông tin thời gian hoặc cả hai, tạo thành dữ liệu không gian, dữ liệu thời gian, và dữ liệu không gian – thời gian tương ứng. Các phương pháp khai phá dữ liệu đặc thù, chẳng hạn như khai phá dữ liệu không gian, khai phá dữ liệu thời gian, khai phá dữ liệu không gian – thời gian, hoặc khai phá mẫu quỹ đạo, cần được phát triển để khai phá các tập dữ liệu như vậy.

Đối với dữ liệu đồ thị và mạng, các ứng dụng khác nhau cũng có thể cần các phương pháp khai phá dữ liệu khác biệt. Ví dụ, mạng xã hội (ví dụ: dữ liệu từ Facebook hoặc LinkedIn), mạng truyền thông máy tính, mạng sinh học, và mạng thông tin (ví dụ: các tác giả liên kết với từ khóa) có thể mang ý nghĩa ngữ nghĩa khác nhau và đòi hỏi các phương pháp khai phá khác nhau.

Ngay cả đối với cùng một tập dữ liệu, việc tìm kiếm các loại mẫu hoặc tri thức khác nhau cũng có thể đòi hỏi các phương pháp khai phá dữ liệu khác nhau. Ví dụ, đối với cùng một tập hợp các chương trình phần mềm (mã nguồn), việc tìm kiếm các module chương trình bị đạo văn hay tìm kiếm các lỗi do copy – paste có thể cần đến những kỹ thuật khai phá dữ liệu khác nhau.

Những loại dữ liệu phong phú và yêu cầu ứng dụng đa dạng đòi hỏi các phương pháp khai phá dữ liệu rất đa dạng. Do đó, khai phá dữ liệu là một lĩnh vực nghiên cứu phong phú và đầy hấp dẫn, với rất nhiều phương pháp mới đang chờ được nghiên cứu và phát triển.

1.3.3 Dữ liệu lưu trữ vs. dữ liệu luồng

Thông thường, khai phá dữ liệu xử lý các tập dữ liệu hữu hạn, được lưu trữ, chẳng hạn như những dữ liệu được lưu trữ trong các kho dữ liệu lớn khác nhau. Tuy nhiên, trong một số ứng dụng như giám sát video hay cảm biến từ xa, dữ liệu có thể được truyền vào một cách động và liên tục, tạo thành các *luồng dữ liệu* vô hạn. Việc khai phá dữ liệu luồng sẽ đòi hỏi các phương pháp khác so với dữ liệu lưu trữ, điều này có thể mở ra một chủ đề nghiên cứu thú vị khác trong lĩnh vực khai phá dữ liệu.

1.4 Khai phá các loại tri thức khác nhau

Các loại mẫu và tri thức khác nhau có thể được phát hiện thông qua khai phá dữ liệu. Nói chung, các nhiệm vụ khai phá dữ liệu có thể được chia thành hai nhóm: **khai phá mô tả** và **khai phá dự đoán**. Khai phá mô tả thực hiện mô tả các thuộc tính của tập dữ liệu quan tâm, trong khi khai phá dự đoán thực hiện quá trình quy nạp trên tập dữ liệu nhằm đưa ra các dự đoán.

Trong mục này, chúng tôi giới thiệu các nhiệm vụ khai phá dữ liệu khác nhau. Các nhiệm vụ này bao gồm: tóm tắt dữ liệu đa chiều (Mục 1.4.1); khai phá các mẫu thường xuyên, các luật kết hợp và mối tương quan (Mục 1.4.2); phân loại và hồi quy (Mục 1.4.3); phân tích cụm (Mục 1.4.4); và phân tích ngoại lệ (Mục 1.4.6). Các

chức năng khai phá dữ liệu khác nhau tạo ra các kết quả khác nhau, thường được gọi là mẫu, mô hình hoặc tri thức. Trong Mục 1.4.7, chúng tôi cũng sẽ giới thiệu về độ thú vị (interestingness) của một mẫu hoặc mô hình. Trong nhiều trường hợp, chỉ những mẫu hoặc mô hình thú vị mới được xem là *tri thức*.

1.4.1 Tóm tắt dữ liệu đa chiều

Đối với người dùng, việc xem qua chi tiết của một tập dữ liệu lớn thường rất mất thời gian. Vì vậy, người dùng mong muốn tự động tóm tắt tập dữ liệu quan tâm và so sánh nó với các tập dữ liệu đối lập ở một số mức độ cao. Miêu tả tóm tắt như vậy của một *tập dữ liệu quan tâm* được gọi là **tóm tắt dữ liệu**. Việc tóm tắt dữ liệu thường được tiến hành trong không gian nhiều chiều. Nếu không gian đa chiều được định nghĩa rõ ràng và được sử dụng thường xuyên, chẳng hạn như danh mục sản phẩm, nhà sản xuất, vị trí hay thời gian, thì một khối lượng dữ liệu khổng lồ có thể được tổng hợp dưới dạng dữ liệu khối nhằm tạo điều kiện cho người dùng thao tác drill – down hoặc roll – up tới không gian tóm tắt chỉ với thao tác nhấp chuột. Kết quả của việc tóm tắt nhiều chiều có thể được trình bày dưới nhiều hình thức khác nhau như **biểu đồ bánh**, **biểu đồ thanh**, **đường cong**, **dữ liệu khối nhiều chiều** và **bảng nhiều chiều**, bao gồm cả các bảng chéo.

Đối với dữ liệu có cấu trúc, các phương pháp tổng hợp đa chiều đã được phát triển để hỗ trợ việc tính toán trước hoặc tính toán trực tuyến của các tổng hợp nhiều chiều sử dụng công nghệ dữ liệu khối, sẽ được bàn luận chi tiết trong Chương 3. Đối với dữ liệu không có cấu trúc, chẳng hạn như văn bản, nhiệm vụ này trở nên đầy thách thức. Chúng tôi sẽ trình bày một số điểm chính về các hướng nghiên cứu tiên tiến này trong chương cuối của cuốn sách.

1.4.2 Khai phá mẫu thường xuyên, luật kết hợp và mối tương quan

Các **mẫu thường xuyên**, như tên gọi đã gợi ý, là những mẫu xuất hiện thường xuyên trong dữ liệu. Có nhiều loại mẫu thường xuyên, bao gồm các tập mục thường xuyên, các chuỗi con thường xuyên (còn được biết đến với tên gọi mẫu tuần tự), và các cấu trúc con thường xuyên. Một *tập mục thường xuyên* thường ám chỉ một tập hợp các mục xuất hiện cùng nhau một cách thường xuyên trong một tập dữ liệu giao dịch — ví dụ, sữa và bánh mì, thường được nhiều khách hàng mua chung tại các cửa hàng tạp hóa. Một chuỗi con xuất hiện thường xuyên, chẳng hạn như

mẫu mà khách hàng có xu hướng mua trước một chiếc laptop, sau đó là túi đựng máy tính và sau cùng là các phụ kiện khác, được gọi là *mẫu tuần tự thường xuyên*. Một cấu trúc con có thể ám chỉ đến các hình thức cấu trúc khác nhau (ví dụ: đồ thị, cây, hoặc lưới hay dàn) có thể được kết hợp với tập mục hoặc chuỗi con. Nếu một cấu trúc con xuất hiện thường xuyên, nó được gọi là *mẫu có cấu trúc thường xuyên*. Việc khai phá các mẫu thường xuyên giúp phát hiện ra các luật kết hợp và mối tương quan thú vị trong dữ liệu.

Ví dụ 2 (Phân tích luật kết hợp)**.** Giả sử, một quản lý cửa hàng trực tuyến muốn biết những mặt hàng nào thường được mua cùng nhau (tức là trong cùng một giao dịch). Một ví dụ về quy tắc như vậy, được khai phá từ cơ sở dữ liệu giao dịch, là:

$$\text{buys}(X, \text{“computer”}) \Rightarrow \text{buys}(X, \text{“webcam”})$$
$$\text{support} = 1\%, \text{confidence} = 50\%,$$

trong đó X là biến đại diện cho khách hàng. Một **độ tin cậy** (confidence) 50% có nghĩa là nếu một khách hàng mua máy tính, có 50% khả năng rằng người đó cũng sẽ mua webcam. Một mức **giá** (support) 1% có nghĩa là 1% tổng số giao dịch được phân tích cho thấy máy tính và webcam được mua cùng nhau. Luật kết hợp này liên quan đến một thuộc tính hoặc mệnh đề duy nhất (ví dụ, *buys*) được lặp lại. Các luật kết hợp chứa một mệnh đề duy nhất được gọi là **luật kết hợp một chiều** (single-dimensional association rule). Bỏ qua ký hiệu mệnh đề, quy tắc có thể được viết đơn giản là “computer ⇒ webcam [1%, 50%]”.
Giả sử, khai phá cùng một cơ sở dữ liệu tạo ra một luật kết hợp khác:

$$\text{age}(X, \text{“20..29”}) \wedge \text{income}(X, \text{“40K..49K”}) \Rightarrow \text{buys}(X, \text{“laptop”})$$
$$\text{support} = 0.5\%, \text{confidence} = 60\%.$$

Quy tắc này cho biết rằng trong tổng số khách hàng được nghiên cứu, 0.5% có độ tuổi từ 20 đến 29 với thu nhập từ 40.000 đến 49.000 USD và đã mua laptop (máy tính xách tay). Có xác suất 60% để một khách hàng trong nhóm độ tuổi và thu nhập này sẽ mua laptop. Lưu ý rằng đây là một luật kết hợp liên quan đến nhiều thuộc tính hoặc mệnh đề (ví dụ, *age*, *income* và *buys*). Theo thuật ngữ được sử dụng trong các cơ sở dữ liệu nhiều chiều, trong đó mỗi thuộc tính được gọi là một chiều, quy tắc trên có thể được gọi là **luật kết hợp nhiều chiều**.

Thông thường, các luật kết hợp sẽ bị loại bỏ nếu chúng không thỏa mãn cả

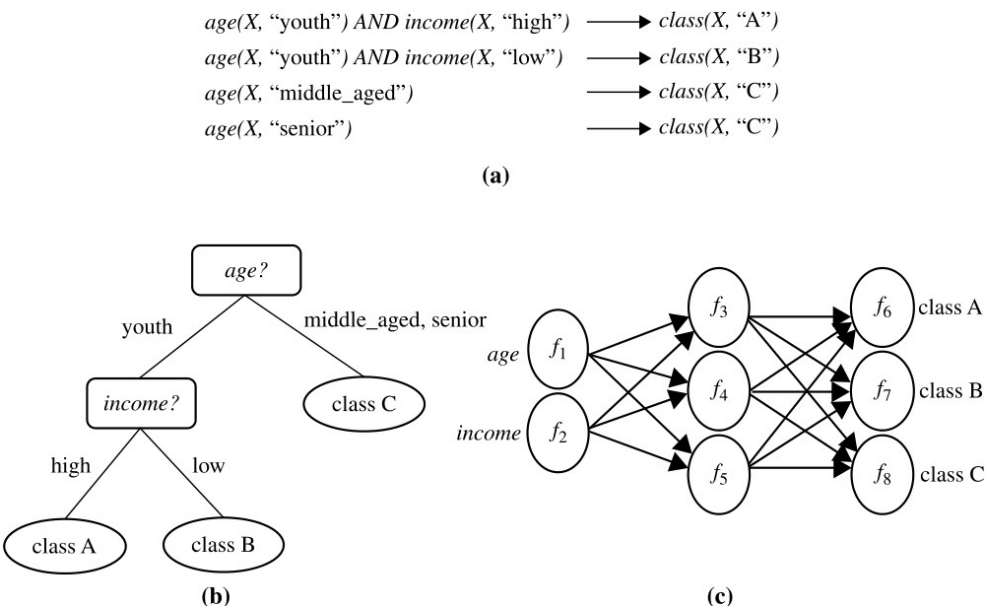
ngưỡng giá tối thiểu và **ngưỡng độ tin cậy tối thiểu**. Các phân tích bổ sung có thể được thực hiện để phát hiện các **tương quan** thống kê thú vị giữa các cặp thuộc tính – giá trị liên quan.

Khai phá tập mục thường xuyên là một hình thức cơ bản của khai phá mẫu thường xuyên. Việc khai phá các tập mục thường xuyên, các luật kết hợp và tương quan sẽ được bàn luận chi tiết trong Chương 4. Việc khai phá các loại mẫu thường xuyên đa dạng, cũng như khai phá mẫu tuần tự và mẫu có cấu trúc, sẽ được trình bày trong Chương 5.

1.4.3 Phân loại và hồi quy cho phân tích dự báo

Phân loại là quá trình tìm kiếm một **mô hình** (hoặc hàm số) mô tả và phân biệt các lớp dữ liệu hoặc các khái niệm. Mô hình này được xây dựng dựa trên phân tích của một tập **dữ liệu huấn luyện** (tức là các đối tượng dữ liệu mà nhãn lớp đã được biết). Mô hình sau đó được sử dụng để dự đoán nhãn lớp của các đối tượng mà nhãn lớp chưa được biết.

Tùy thuộc vào các phương pháp phân loại, mô hình thu được có thể có nhiều hình thức khác nhau, chẳng hạn như tập hợp các *quy tắc phân loại* (ví dụ: *quy tắc IF – THEN*), *cây quyết định*, *công thức toán học*, hoặc một *mạng nơron* đã được huấn luyện (xem Hình 1.2).



Hình 1.2: Một mô hình phân loại có thể được biểu diễn dưới nhiều hình thức: (a) quy tắc IF – THEN, (b) cây quyết định, hoặc (c) mạng nơron.

Cây quyết định là một cấu trúc dạng sơ đồ luồng, trong đó mỗi nút biểu thị một phép kiểm tra trên giá trị thuộc tính, mỗi nhánh đại diện cho một kết quả của phép kiểm tra đó, và các lá của cây biểu thị các lớp hoặc phân phối của các lớp. Cây quyết định có thể được dễ dàng chuyển đổi thành các quy tắc phân loại. Một **mạng nơron**, khi được sử dụng cho phân loại, thường là một tập hợp các đơn vị xử lý giống như nơron với các kết nối có trọng số giữa các đơn vị. Ngoài ra, còn có nhiều phương pháp khác để xây dựng mô hình phân loại, chẳng hạn như phân loại Bayes đơn giản, máy véctơ hỗ trợ (SVM) và phân loại $k-$láng giềng gần nhất.

Trong khi phân loại dự đoán các nhãn rời rạc (không có thứ tự), **hồi quy** mô hình hóa các hàm có giá trị liên tục. Nói cách khác, hồi quy được sử dụng để dự đoán các *giá trị số* bị thiếu hoặc không có sẵn thay vì các nhãn (rời rạc) của lớp. Thuật ngữ "*dự đoán*" đề cập đến cả dự đoán số và dự đoán nhãn lớp. **Phân tích hồi quy** là một phương pháp thống kê được sử dụng phổ biến nhất cho dự đoán số, mặc dù có nhiều phương pháp khác cũng tồn tại. Hồi quy cũng bao gồm việc xác định *xu hướng* phân phối dựa trên dữ liệu có sẵn.

Phân loại và hồi quy có thể cần được thực hiện sau bước **lựa chọn đặc trưng** hoặc **phân tích mức độ liên quan**, nhằm xác định các thuộc tính (thường được gọi là các *đặc trưng*) có liên quan đáng kể đến quá trình phân loại và hồi quy. Những thuộc tính này sẽ được chọn cho quá trình phân loại và hồi quy, trong khi các thuộc tính không liên quan có thể được loại bỏ khỏi quá trình phân tích.

Ví dụ 3 (Phân loại và hồi quy)**.** Giả sử, một quản lý bán hàng của cửa hàng trực tuyến muốn phân loại một tập hợp lớn các mặt hàng trong cửa hàng dựa trên ba loại phản hồi đối với một chiến dịch bán hàng: *phản hồi tốt*, *phản hồi vừa phải* và *không có phản hồi*. Bạn muốn xây dựng một mô hình cho mỗi ba lớp này dựa trên các đặc trưng mô tả của các mặt hàng, chẳng hạn như *giá*, *thương hiệu*, *nơi sản xuất*, *loại* và *danh mục*. Phân loại thu được cần phân biệt tối đa giữa các lớp, tạo nên một bức tranh có tổ chức của tập dữ liệu.

Giả sử kết quả phân loại được biểu diễn dưới dạng một cây quyết định. Ví dụ, cây quyết định có thể xác định *giá* là yếu tố quan trọng đầu tiên giúp phân biệt ba lớp này. Các đặc trưng khác hỗ trợ thêm trong việc phân biệt các đối tượng thuộc mỗi lớp bao gồm *thương hiệu* và *nơi sản xuất*. Cây quyết định như vậy có thể giúp quản lý hiểu được tác động của chiến dịch bán hàng đã triển khai và từ đó thiết kế ra một chiến dịch hiệu quả hơn trong tương lai.

Giả sử, thay vào đó, bạn không muốn dự đoán các nhãn phản hồi rời rạc cho mỗi mặt hàng trong cửa hàng mà bạn muốn dự đoán doanh thu mà mỗi mặt hàng sẽ

tạo ra trong một đợt giảm giá sắp tới, dựa trên dữ liệu bán hàng trước đó. Đây là một ví dụ về phân tích hồi quy bởi vì mô hình hồi quy được xây dựng sẽ dự đoán một hàm số liên tục (hoặc giá trị có thứ tự).

Chương 6 và 7 bàn sâu hơn về phân loại. Phân tích hồi quy chỉ được đề cập sơ qua trong các chương này vì nó thường được giới thiệu trong các khóa học thống kê. Các nguồn tham khảo thêm được liệt kê trong ghi chú tài liệu tham khảo.

1.4.4 Phân tích cụm

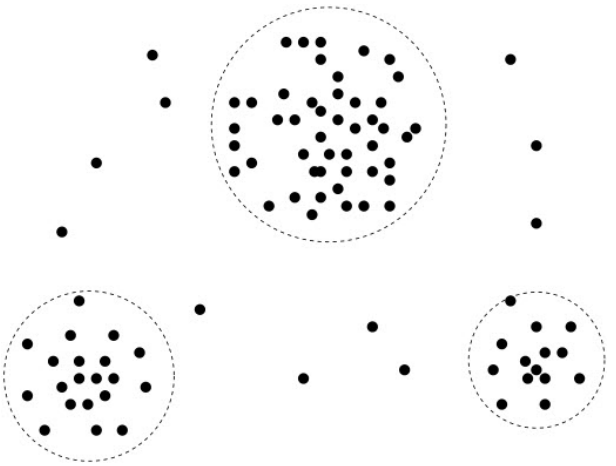
Khác với phân loại và hồi quy, vốn phân tích các tập dữ liệu có nhãn (dữ liệu huấn luyện), **phân tích cụm** nhóm các đối tượng dữ liệu mà không dựa vào nhãn lớp. Trong nhiều trường hợp, dữ liệu có nhãn có thể không tồn tại ngay từ ban đầu. Phân cụm có thể được sử dụng để tạo ra các nhãn lớp cho một nhóm dữ liệu. Các đối tượng được nhóm lại dựa trên nguyên tắc *tối đa hóa độ tương đồng trong cùng một lớp và tối thiểu hóa độ tương đồng giữa các lớp*. Nói cách khác, các cụm đối tượng được hình thành sao cho các đối tượng trong cùng một cụm có độ tương đồng cao với nhau, nhưng lại khá khác biệt so với các đối tượng ở các cụm khác. Mỗi cụm được hình thành như vậy có thể được xem như một lớp đối tượng, từ đó có thể trích xuất ra các quy tắc. Phân cụm cũng hỗ trợ việc hình thành **phân loại học** (taxonomy), tức là việc tổ chức các quan sát thành một hệ thống phân cấp các lớp gom nhóm các sự kiện tương đồng lại với nhau.

Ví dụ 4 (Phân tích cụm)**.** Phân tích cụm có thể được áp dụng trên dữ liệu khách hàng của cửa hàng trực tuyến để xác định các nhóm khách hàng đồng nhất. Các cụm này có thể đại diện cho từng nhóm mục tiêu riêng biệt cho chiến dịch tiếp thị. Hình 1.3 cho thấy một đồ thị 2-D của khách hàng theo vị trí của họ trong một thành phố. Ba cụm điểm dữ liệu rõ ràng hiện ra.

Phân tích cụm là chủ đề của Chương 8 và 9.

1.4.5 Học sâu

Đối với nhiều nhiệm vụ khai phá dữ liệu, chẳng hạn như phân loại và phân cụm, một bước quan trọng thường nằm ở việc tìm ra "đặc trưng tốt", tức là cách biểu diễn dưới dạng véctơ cho mỗi bộ dữ liệu đầu vào. Ví dụ, để dự đoán khả năng bùng phát dịch bệnh theo khu vực, có thể đã thu thập một số lượng lớn các đặc



Hình 1.3: Một biểu đồ 2 chiều của dữ liệu khách hàng theo vị trí của khách hàng trong thành phố, hiển thị ba cụm dữ liệu.

trưng từ dữ liệu giám sát sức khỏe, bao gồm số ca dương tính hàng ngày, số ca xét nghiệm hàng ngày, số ca nhập viện hàng ngày, v.v. Theo truyền thống, bước này (gọi là kỹ thuật lấy đặc trưng – feature engineering) thường phụ thuộc nặng nề vào kiến thức chuyên ngành. Các kỹ thuật học sâu cung cấp một cách tự động để thực hiện kỹ thuật lấy đặc trưng, có khả năng tạo ra các đặc trưng có ý nghĩa ngữ nghĩa (ví dụ: tỷ lệ dương tính hàng tuần) từ các đặc trưng đầu vào ban đầu. Các đặc trưng được tạo ra thường cải thiện đáng kể hiệu năng của quá trình khai phá (ví dụ: độ chính xác phân loại).

Học sâu dựa trên các *mạng nơron*. Một mạng nơron là tập hợp các đơn vị đầu vào – đầu ra được kết nối với nhau, trong đó mỗi kết nối được gán một trọng số. Trong quá trình học, mạng sẽ tự điều chỉnh các trọng số sao cho có thể dự đoán đúng giá trị mục tiêu (ví dụ: nhãn lớp) của các bộ dữ liệu đầu vào. Thuật toán cốt lõi để học các trọng số này được gọi là *lan truyền ngược* (backpropagation), nó tìm kiếm một tập hợp các trọng số và giá trị chệch sao cho mô hình hóa dữ liệu đạt được mức độ lỗi (loss function) tối thiểu giữa dự đoán của mạng và giá trị mục tiêu thực tế của các bộ dữ liệu. Nhiều dạng kiến trúc khác nhau của mạng nơron đã được phát triển, bao gồm mạng nơron truyền thẳng (feed-forward neural networks), mạng nơron tích chập (convolutional neural networks), mạng nơron hồi tiếp (recurrent neural networks), mạng nơron đồ thị (graph neural networks) và nhiều dạng khác.

Học sâu có ứng dụng rộng rãi trong thị giác máy tính, xử lý ngôn ngữ tự nhiên, dịch máy, phân tích mạng xã hội, và nhiều lĩnh vực khác. Nó đã được sử dụng trong

nhiều nhiệm vụ khai phá dữ liệu, bao gồm phân loại, phân cụm, phát hiện điểm bất thường và học tăng cường.

Chủ đề về học sâu được trình bày chi tiết trong Chương 10.

1.4.6 Phân tích ngoại lệ

Một tập dữ liệu có thể chứa các đối tượng không tuân theo hành vi hay mô hình chung của dữ liệu, gọi là **ngoại lệ**, hay điểm bất thường. Nhiều phương pháp khai phá dữ liệu thường loại bỏ các ngoại lệ như là nhiễu. Tuy nhiên, trong một số ứng dụng (ví dụ: phát hiện gian lận), các sự kiện hiếm gặp lại có thể thú vị hơn so với những sự kiện xảy ra đều đặn. Việc phân tích các dữ liệu bất thường được gọi là **phân tích ngoại lệ** hoặc **khai phá điểm bất thường**.

Các ngoại lệ có thể được phát hiện bằng cách sử dụng các kiểm định thống kê dựa trên giả định về phân phối hoặc mô hình xác suất của dữ liệu, hoặc sử dụng các phép đo khoảng cách, trong đó các đối tượng nằm xa bất kỳ cụm nào được xem là điểm bất thường. Thay vì sử dụng các phép đo thống kê hay khoảng cách, các phương pháp dựa trên mật độ có thể nhận dạng điểm bất thường trong một vùng cục bộ, mặc dù chúng có vẻ bình thường khi nhìn theo phân phối thống kê toàn cục.

Ví dụ 5 (Phân tích ngoại lệ)**.** Phân tích ngoại lệ có thể phát hiện ra việc sử dụng thẻ tín dụng gian lận bằng cách nhận diện các giao dịch mua với số tiền lớn bất thường đối với một số tài khoản cụ thể, so với các khoản phí thông thường mà cùng tài khoản đó thường gặp. Các giá trị bất thường cũng có thể được phát hiện dựa trên vị trí và loại giao dịch, hoặc tần suất giao dịch.

Phân tích điểm bất thường được bàn luận chi tiết trong Chương 11.

1.4.7 Phải chăng tất cả các kết quả khai phá đều thú vị?

Khai phá dữ liệu có khả năng tạo ra rất nhiều kết quả. Một câu hỏi được đặt ra là: “Có phải tất cả các kết quả khai phá đều thú vị không?”

Đây là một câu hỏi hay. Mỗi loại chức năng khai phá dữ liệu đều có các chỉ số riêng để đánh giá chất lượng của quá trình khai phá. Tuy nhiên, có một số triết lý và nguyên tắc chung.

Lấy khai thác mẫu làm ví dụ, quá trình này có thể tạo ra hàng nghìn, thậm chí hàng triệu mẫu hoặc quy tắc. Bạn có thể tự hỏi, “*Điều gì làm cho một mẫu trở nên*

thú vị? Liệu một hệ thống khai phá dữ liệu có thể tạo ra tất cả các mẫu thú vị, hoặc chỉ các mẫu thú vị?"

Để trả lời câu hỏi đầu tiên, một mẫu được xem là **thú vị** nếu nó thỏa mãn các tiêu chí sau: (1) *dễ hiểu* đối với con người, (2) *có hiệu lực* trên dữ liệu mới hoặc dữ liệu thử nghiệm với một mức độ *chắc chắn nhất định*, (3) có tiềm năng *hữu ích*, và (4) *mới mẻ, độc đáo*. Một mẫu cũng được coi là thú vị nếu nó khẳng định một giả thuyết mà người dùng *muốn xác nhận*.

Có một số **chỉ số khách quan để đo lường độ thú vị của mẫu**, dựa trên cấu trúc của các mẫu được phát hiện và các thống kê liên quan. Một chỉ số khách quan đối với các luật kết hợp dạng $X \Rightarrow Y$ là quy tắc **giá** (support), thể hiện phần trăm các giao dịch trong cơ sở dữ liệu giao dịch mà quy tắc cho trước thỏa mãn. Điều này được hiểu là xác suất $P(X \cup Y)$, trong đó $X \cup Y$ cho biết một giao dịch chứa cả X và Y, tức là hợp của các tập mục X và Y. Một chỉ số khách quan khác đối với luật kết hợp là **độ tin cậy** (confidence), đánh giá mức độ chắc chắn của mối liên kết được phát hiện. Đây được hiểu là xác suất có điều kiện $P\left(Y \mid X\right)$, tức là xác suất một giao dịch nếu chứa X thì cũng chứa Y. Một cách chính thức, giá và độ tin cậy được định nghĩa như sau:

$$\text{support}\,(X \Rightarrow Y) = P(X \cup Y), \tag{1.1}$$

$$\text{confidence}\,(X \Rightarrow Y) = P\left(Y \mid X\right). \tag{1.2}$$

Nói chung, mỗi chỉ số đo độ thú vị sẽ có một ngưỡng nhất định, được người dùng kiểm soát. Ví dụ, các quy tắc không đạt được ngưỡng độ tin cậy, chẳng hạn 50%, có thể được coi là không thú vị. Các quy tắc dưới ngưỡng có khả năng phản ánh nhiễu, ngoại lệ hoặc các trường hợp thiểu số và có lẽ ít giá trị hơn.

Cũng có các chỉ số khách quan khác. Ví dụ, có người có thể muốn tập hợp các mục trong một luật kết hợp có mối tương quan mạnh mẽ. Chúng tôi sẽ bàn về những chỉ số như vậy trong chương tương ứng.

Mặc dù các chỉ số khách quan giúp xác định các mẫu thú vị, nhưng chúng thường không đủ nếu không kết hợp với các chỉ số chủ quan phản ánh nhu cầu và sở thích cụ thể của người dùng. Ví dụ, các mẫu mô tả đặc điểm của khách hàng thường xuyên mua sắm trực tuyến có thể rất thú vị đối với quản lý tiếp thị, nhưng lại ít thu hút đối với những nhà phân tích khác nghiên cứu cùng một cơ sở dữ liệu về hiệu suất nhân viên. Hơn nữa, nhiều mẫu mà theo tiêu chuẩn khách quan được xem là thú vị có thể đại diện cho lẽ thường và do đó, thực sự không có nhiều giá trị.

Các **chỉ số độ thú vị chủ quan** dựa trên niềm tin của người dùng đối với dữ liệu. Những chỉ số này coi một mẫu là thú vị nếu mẫu đó **bất ngờ** (trái ngược với niềm tin của người dùng) hoặc mang lại thông tin chiến lược mà người dùng có thể ứng dụng. Trong trường hợp sau, các mẫu như vậy được gọi là có tính **khả thi hành động** (actionable). Ví dụ, các mẫu như "một trận động đất lớn thường xuất hiện sau một cụm các trận động đất nhỏ" có thể có tính khả thi hành động cao nếu người dùng có thể ứng dụng thông tin đó để cứu sống người. Các mẫu **được dự đoán** có thể vẫn được xem là thú vị nếu chúng xác nhận một giả thuyết mà người dùng muốn kiểm chứng hoặc phù hợp với cảm nhận ban đầu của người dùng.

Câu hỏi thứ hai — "*Liệu một hệ thống khai phá dữ liệu có thể tạo ra tất cả các mẫu thú vị?*" — liên quan đến **tính đầy đủ** của một thuật toán khai phá. Thường thì không thực tế và không hiệu quả khi một hệ thống khai phá mẫu cố gắng tạo ra tất cả các mẫu có thể có vì số lượng chúng có thể rất lớn. Tuy nhiên, cũng có thể lo lắng rằng có thể bỏ sót những mẫu quan trọng nếu hệ thống dừng lại quá sớm. Để giải quyết tình huống này, các ràng buộc do người dùng cung cấp và các chỉ số độ thú vị nên được sử dụng để tập trung việc tìm kiếm. Với các chỉ số độ thú vị được xác định rõ và các ràng buộc do người dùng cung cấp, việc đảm bảo tính đầy đủ của khai phá mẫu là hoàn toàn khả thi. Các phương pháp liên quan được trình bày chi tiết trong Chương 4.

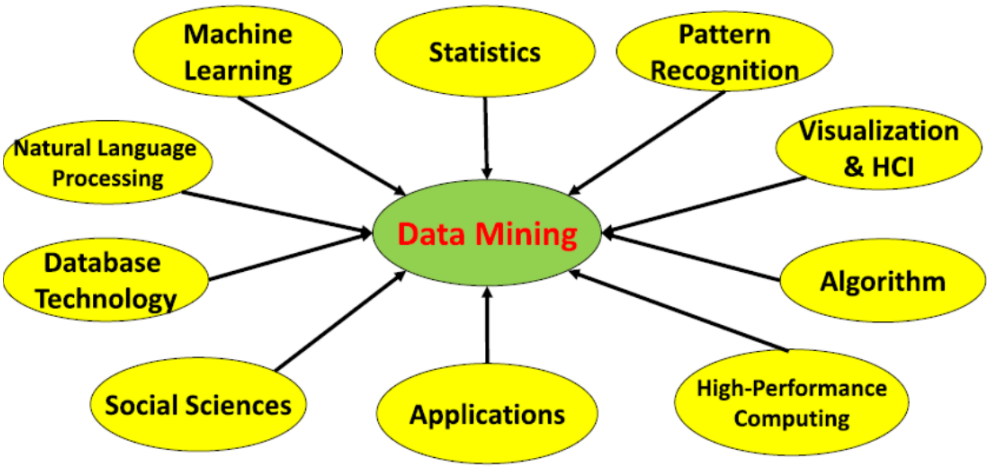
Cuối cùng, câu hỏi thứ ba — "*Liệu một hệ thống khai phá dữ liệu có thể tạo ra chỉ các mẫu thú vị?*" — là một bài toán tối ưu trong khai phá dữ liệu. Người dùng rất mong muốn một hệ thống khai phá dữ liệu chỉ tạo ra những mẫu thú vị. Điều này sẽ hiệu quả cho cả hệ thống và người dùng bởi vì hệ thống có thể tốn ít thời gian hơn để tạo ra một số lượng mẫu ít nhưng chất lượng, trong khi người dùng sẽ không phải sàng lọc qua một lượng lớn mẫu để xác định những mẫu thực sự thú vị. Khai phá mẫu dựa trên ràng buộc, được mô tả trong Chương 5, là một ví dụ điển hình theo hướng này.

Các phương pháp đánh giá chất lượng hoặc độ thú vị của kết quả khai phá dữ liệu, và cách sử dụng chúng để cải thiện hiệu quả khai phá dữ liệu, được bàn luận xuyên suốt cuốn sách.

1.5 Khai phá dữ liệu: sự giao thoa của nhiều lĩnh vực

Là một lĩnh vực nghiên cứu các phương pháp hiệu quả và tối ưu để khám phá các mẫu và tri thức từ nhiều loại tập dữ liệu khổng lồ phục vụ cho nhiều ứng dụng

khác nhau, khai phá dữ liệu tự nhiên đóng vai trò là sự giao thoa của nhiều ngành, bao gồm học máy, thống kê, nhận dạng mẫu, xử lý ngôn ngữ tự nhiên, công nghệ cơ sở dữ liệu, trực quan hóa và tương tác người – máy (human computer interaction, HCI), thuật toán, tính toán hiệu năng cao, khoa học xã hội và nhiều lĩnh vực ứng dụng khác (Hình 1.4). Tính liên ngành trong nghiên cứu và phát triển khai phá dữ liệu đóng góp đáng kể vào sự thành công của lĩnh vực này cũng như vào phạm vi ứng dụng rộng lớn của nó. Mặt khác, khai phá dữ liệu không chỉ được nuôi dưỡng từ tri thức và sự phát triển của các ngành này mà còn tác động mạnh mẽ đến sự phát triển của chúng trong những năm gần đây thông qua các nghiên cứu, phát triển và ứng dụng chuyên biệt trên nhiều loại dữ liệu lớn khác nhau. Trong mục này, chúng ta sẽ thảo luận về một số ngành có ảnh hưởng mạnh mẽ và có sự tương tác chặt chẽ với nghiên cứu, phát triển và ứng dụng khai phá dữ liệu.



Hình 1.4: Khai phá dữ liệu: sự giao thoa của nhiều lĩnh vực.

1.5.1 Thống kê và khai phá dữ liệu

Thống kê nghiên cứu việc thu thập, phân tích, diễn giải hoặc giải thích, và trình bày dữ liệu. Khai phá dữ liệu có mối liên hệ chặt chẽ với thống kê.

Một **mô hình thống kê** là một tập hợp các hàm toán học mô tả hành vi của các đối tượng trong một lớp mục tiêu dưới dạng các biến ngẫu nhiên và các phân phối xác suất liên quan. Các mô hình thống kê được sử dụng rộng rãi để mô hình hóa dữ liệu và các lớp dữ liệu. Chẳng hạn, trong các tác vụ khai phá dữ liệu như đặc trưng hóa dữ liệu và phân loại, có thể xây dựng các mô hình thống kê cho các lớp mục tiêu. Nói cách khác, các mô hình thống kê như vậy có thể là kết quả của một

tác vụ khai phá dữ liệu. Ngược lại, các tác vụ khai phá dữ liệu cũng có thể được xây dựng dựa trên các mô hình thống kê. Ví dụ, ta có thể sử dụng thống kê để mô hình hóa nhiễu và các giá trị bị thiếu trong dữ liệu. Vì vậy, khi khai phá các mẫu trong một tập dữ liệu lớn, quá trình khai phá dữ liệu có thể sử dụng mô hình thống kê để xác định và xử lý các giá trị nhiễu hoặc bị thiếu trong dữ liệu.

Nghiên cứu thống kê phát triển các công cụ để dự đoán và ước lượng dựa trên dữ liệu và các mô hình thống kê. Các phương pháp thống kê có thể được sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu. Các **mô tả thống kê** cơ bản của dữ liệu sẽ được giới thiệu trong Chương 2. Thống kê hữu ích trong việc khai phá nhiều mẫu khác nhau từ dữ liệu và giúp hiểu rõ các cơ chế cơ bản tạo ra và ảnh hưởng đến các mẫu này. **Thống kê suy diễn** (còn gọi là **thống kê dự báo**) mô hình hóa dữ liệu theo cách tính đến tính ngẫu nhiên và sự không chắc chắn trong quan sát, từ đó đưa ra suy luận về quá trình hoặc quần thể đang được nghiên cứu.

Các phương pháp thống kê cũng có thể được sử dụng để kiểm định kết quả khai phá dữ liệu. Ví dụ, sau khi một mô hình phân loại hoặc dự đoán được khai phá, mô hình này cần được kiểm định thông qua kiểm định giả thuyết thống kê. Một **kiểm định giả thuyết thống kê** (đôi khi được gọi là *phân tích dữ liệu chứng thực*) đưa ra quyết định thống kê dựa trên dữ liệu thực nghiệm. Một kết quả được gọi là có *ý nghĩa thống kê* nếu khả năng nó xảy ra một cách ngẫu nhiên là rất thấp. Nếu mô hình phân loại hoặc dự đoán là đúng, thì các thống kê mô tả của mô hình sẽ củng cố tính hợp lý của mô hình.

Việc áp dụng các phương pháp thống kê trong khai phá dữ liệu không hề đơn giản. Một thách thức lớn là làm thế nào để mở rộng một phương pháp thống kê cho một tập dữ liệu lớn. Nhiều phương pháp thống kê có độ phức tạp tính toán cao. Khi áp dụng các phương pháp này trên các tập dữ liệu lớn được phân tán trên nhiều địa điểm logic hoặc vật lý, các thuật toán cần được thiết kế và điều chỉnh cẩn thận để giảm chi phí tính toán. Thách thức này trở nên khó khăn hơn đối với các ứng dụng trực tuyến, chẳng hạn như gợi ý truy vấn trực tuyến trong các công cụ tìm kiếm, nơi mà khai phá dữ liệu phải liên tục xử lý các luồng dữ liệu thời gian thực với tốc độ cao.

Nghiên cứu khai phá dữ liệu đã phát triển nhiều giải pháp hiệu quả và có thể mở rộng quy mô để phân tích các tập dữ liệu lớn và luồng dữ liệu. Hơn nữa, các loại dữ liệu khác nhau và các ứng dụng khác nhau có thể yêu cầu những phương pháp phân tích rất khác nhau. Nhiều giải pháp hiệu quả đã được đề xuất và kiểm định, từ đó dẫn đến sự phát triển của nhiều phương pháp mới và có thể mở rộng trong phân tích thống kê dựa trên khai phá dữ liệu.

1.5.2 Học máy và khai phá dữ liệu

Học máy nghiên cứu cách máy tính có thể học (hoặc cải thiện hiệu suất của chúng) dựa trên dữ liệu. Đây là một lĩnh vực phát triển nhanh chóng, với nhiều phương pháp và ứng dụng mới được phát triển trong những năm gần đây, từ máy véctơ hỗ trợ (SVM) đến mô hình đồ thị xác suất và học sâu, những chủ đề sẽ được đề cập trong cuốn sách này.

Nhìn chung, học máy giải quyết hai bài toán kinh điển: *học có giám sát* và *học không giám sát*.

- **Học có giám sát**: Một ví dụ kinh điển của học có giám sát là phân loại. Việc giám sát trong quá trình học đến từ các ví dụ được gán nhãn trong tập dữ liệu huấn luyện. Ví dụ, để tự động nhận dạng mã bưu điện viết tay trên thư từ, hệ thống học sẽ sử dụng một tập hợp các hình ảnh mã bưu điện viết tay và các bản dịch của chúng thành dạng có thể đọc được bằng máy như các ví dụ huấn luyện, sau đó học (tức là tính toán) một mô hình phân loại.

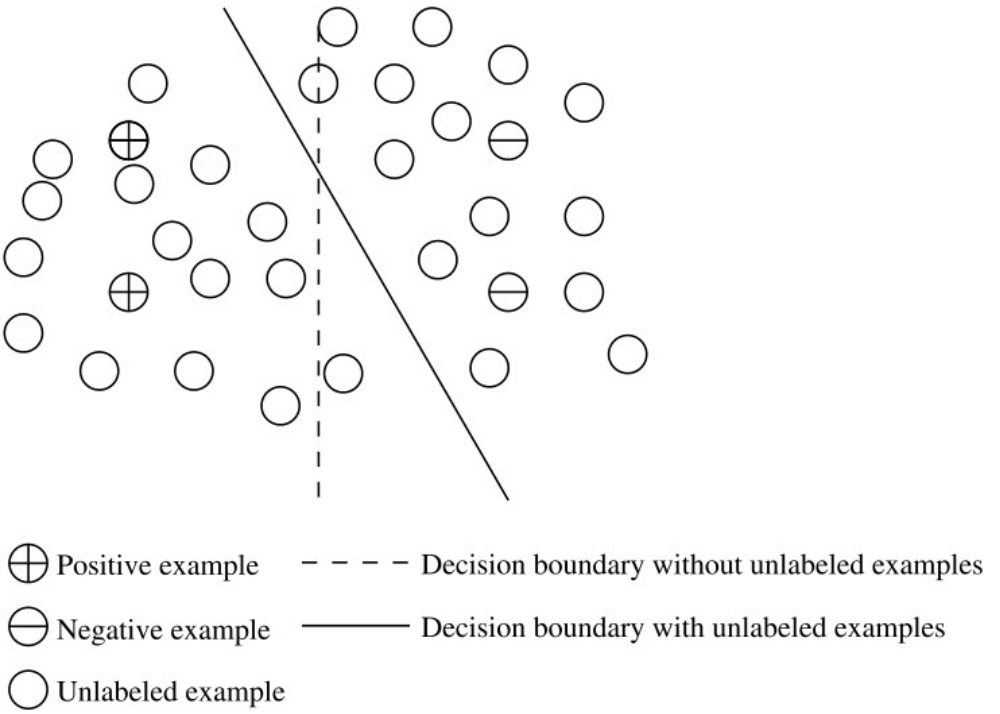
- **Học không giám sát**: Một ví dụ kinh điển của học không giám sát là phân cụm. Quá trình học này không có giám sát vì các ví dụ đầu vào không có nhãn lớp. Thông thường, ta có thể sử dụng phân cụm để khám phá các nhóm trong dữ liệu. Ví dụ, một phương pháp học không giám sát có thể nhận một tập hợp hình ảnh của các chữ số viết tay làm đầu vào. Giả sử nó tìm thấy 10 cụm dữ liệu, có khả năng các cụm này tương ứng với 10 chữ số từ 0 đến 9. Tuy nhiên, vì dữ liệu huấn luyện không có nhãn, mô hình học được không thể cho biết ý nghĩa ngữ nghĩa của các cụm tìm thấy.

Xét trên hai bài toán cơ bản này, khai phá dữ liệu và học máy có nhiều điểm tương đồng. Tuy nhiên, khai phá dữ liệu khác với học máy ở một số khía cạnh quan trọng.

Thứ nhất, ngay cả trên các tác vụ tương tự như phân loại và phân cụm, khai phá dữ liệu thường làm việc với các tập dữ liệu rất lớn, hoặc thậm chí là các luồng dữ liệu vô hạn. Khả năng mở rộng là một mối quan tâm quan trọng, và nhiều thuật toán khai phá dữ liệu hiệu quả và có thể mở rộng, cũng như các thuật toán khai phá dữ liệu trên luồng dữ liệu, đã được phát triển để thực hiện các nhiệm vụ này.

Thứ hai, trong nhiều bài toán khai phá dữ liệu, các tập dữ liệu thường rất lớn, nhưng dữ liệu huấn luyện có thể vẫn khá nhỏ do chi phí gán nhãn dữ liệu chất lượng cao bởi các chuyên gia là rất tốn kém. Do đó, khai phá dữ liệu phải tập trung vào việc phát triển các *phương pháp học có giám sát yếu*. Những phương pháp

này bao gồm *học bán giám sát* với một tập nhỏ dữ liệu được gán nhãn nhưng một tập lớn dữ liệu chưa được gán nhãn (như minh họa trong Hình 1.5), *tích hợp hoặc kết hợp nhiều mô hình yếu* từ những nguồn không chuyên (ví dụ, các mô hình thu được thông qua huy động cộng đồng), *giám sát từ xa* bằng cách sử dụng các cơ sở tri thức phổ biến nhưng có liên quan xa đến vấn đề cần giải quyết (ví dụ: Wikipedia, DBPedia), *học chủ động* bằng cách chọn lựa cẩn thận các ví dụ để yêu cầu chuyên gia con người gán nhãn, hoặc *học chuyển giao* bằng cách tích hợp các mô hình học được từ các lĩnh vực bài toán tương tự. Khai phá dữ liệu đã mở rộng các phương pháp học có giám sát yếu như vậy để xây dựng các mô hình phân loại chất lượng trên các tập dữ liệu lớn với một tập dữ liệu huấn luyện chất lượng cao rất hạn chế.



Hình 1.5: Học bán giám sát.

Thứ ba, các phương pháp học máy có thể không xử lý được nhiều loại bài toán khám phá tri thức trên dữ liệu lớn. Ngược lại, khai phá dữ liệu, khi phát triển các giải pháp hiệu quả cho các bài toán ứng dụng cụ thể, đi sâu vào từng lĩnh vực bài toán và mở rộng phạm vi nghiên cứu vượt xa phạm vi của học máy. Ví dụ, nhiều bài toán ứng dụng như phân tích dữ liệu giao dịch kinh doanh, phân tích trình tự thực thi chương trình phần mềm, và phân tích cấu trúc hóa học và sinh học cần các phương pháp hiệu quả trong khai phá mẫu phổ biến, mẫu trình tự và mẫu có

cấu trúc. Nghiên cứu khai phá dữ liệu đã tạo ra nhiều phương pháp khai phá hiệu quả, đa dạng và có thể mở rộng cho các nhiệm vụ này. Một ví dụ khác là phân tích mạng xã hội và mạng thông tin quy mô lớn, đặt ra nhiều bài toán thách thức mà có thể không phù hợp với phạm vi điển hình của nhiều phương pháp học máy do sự tương tác thông tin giữa các liên kết và nút trong mạng. Khai phá dữ liệu đã phát triển nhiều giải pháp thú vị cho các bài toán như vậy.

Từ góc nhìn này, khai phá dữ liệu và học máy là hai lĩnh vực khác biệt nhưng có liên hệ mật thiết với nhau. Khai phá dữ liệu đi sâu vào các lĩnh vực ứng dụng cụ thể, tập trung vào dữ liệu quy mô lớn, không giới hạn trong một phương pháp giải quyết bài toán đơn lẻ, và phát triển các giải pháp cụ thể (đôi khi khá mới lạ), hiệu quả và có thể mở rộng cho nhiều bài toán ứng dụng thách thức. Đây là một lĩnh vực nghiên cứu trẻ, rộng lớn và đầy hứa hẹn cho nhiều nhà nghiên cứu và thực hành tham gia nghiên cứu và ứng dụng.

1.5.3 Công nghệ cơ sở dữ liệu và khai phá dữ liệu

Nghiên cứu về hệ thống cơ sở dữ liệu tập trung vào việc tạo lập, bảo trì và sử dụng cơ sở dữ liệu cho các tổ chức và người dùng cuối. Đặc biệt, các nhà nghiên cứu về hệ thống cơ sở dữ liệu đã thiết lập những nguyên tắc được công nhận rộng rãi trong các lĩnh vực như mô hình dữ liệu, ngôn ngữ truy vấn, xử lý và tối ưu truy vấn, lưu trữ dữ liệu và phương pháp lập chỉ mục. Công nghệ cơ sở dữ liệu nổi tiếng với khả năng mở rộng trong xử lý các tập dữ liệu rất lớn và có cấu trúc tương đối tốt.

Nhiều tác vụ khai phá dữ liệu cần xử lý các tập dữ liệu lớn hoặc thậm chí là luồng dữ liệu nhanh theo thời gian thực. Khai phá dữ liệu có thể tận dụng công nghệ cơ sở dữ liệu có khả năng mở rộng để đạt hiệu suất cao và khả năng mở rộng trên các tập dữ liệu lớn. Hơn nữa, các tác vụ khai phá dữ liệu có thể được sử dụng để mở rộng khả năng của các hệ thống cơ sở dữ liệu hiện có nhằm đáp ứng các yêu cầu phân tích dữ liệu phức tạp của người dùng.

Các hệ thống cơ sở dữ liệu hiện đại đã tích hợp khả năng phân tích dữ liệu một cách có hệ thống thông qua kho dữ liệu và các công cụ khai phá dữ liệu. **Kho dữ liệu** tổng hợp dữ liệu từ nhiều nguồn và nhiều khoảng thời gian khác nhau. Nó hợp nhất dữ liệu trong không gian nhiều chiều để tạo thành các dữ liệu khối được lưu trữ một phần. Mô hình dữ liệu khối không chỉ hỗ trợ xử lý phân tích trực tuyến (OLAP) trong các cơ sở dữ liệu nhiều chiều mà còn thúc đẩy *khai phá dữ liệu nhiều chiều*, nội dung sẽ được thảo luận chi tiết trong các chương tiếp theo.

1.5.4 Khai phá dữ liệu và khoa học dữ liệu

Với lượng dữ liệu khổng lồ xuất hiện trong hầu hết các lĩnh vực và các ứng dụng khác nhau, thuật ngữ dữ liệu lớn và khoa học dữ liệu đã trở thành những từ khóa phổ biến trong những năm gần đây. **Dữ liệu lớn** thường chỉ khối lượng dữ liệu rất lớn, bao gồm cả dữ liệu có cấu trúc và không có cấu trúc dưới nhiều dạng khác nhau. Trong khi đó, **khoa học dữ liệu** là một lĩnh vực liên ngành sử dụng các phương pháp, quy trình, thuật toán và hệ thống khoa học để trích xuất tri thức và thông tin từ dữ liệu lớn thuộc nhiều dạng khác nhau. Rõ ràng, khai phá dữ liệu đóng một vai trò thiết yếu trong khoa học dữ liệu.

Đối với hầu hết mọi người, khoa học dữ liệu là một khái niệm thống nhất các lĩnh vực thống kê, học máy, khai phá dữ liệu và các phương pháp liên quan nhằm mục đích hiểu và phân tích dữ liệu lớn. Lĩnh vực này sử dụng các kỹ thuật và lý thuyết rút ra từ nhiều lĩnh vực trong toán học, thống kê, khoa học thông tin và khoa học máy tính. Đối với nhiều người trong ngành công nghiệp, thuật ngữ "khoa học dữ liệu" thường được sử dụng để chỉ phân tích kinh doanh, trí tuệ kinh doanh, mô hình dự đoán, hoặc bất kỳ ứng dụng nào có ý nghĩa trong việc khai thác dữ liệu. Một số người còn coi đó là một thuật ngữ hấp dẫn được sử dụng để "tái định nghĩa" các lĩnh vực như thống kê, khai phá dữ liệu, học máy hoặc bất kỳ dạng phân tích dữ liệu nào khác. Hiện nay, vẫn chưa có một định nghĩa thống nhất hay một chương trình đào tạo tiêu chuẩn cho các ngành khoa học dữ liệu tại các trường đại học. Tuy nhiên, hầu hết các trường đại học đều coi các kiến thức nền tảng từ thống kê, học máy, khai phá dữ liệu, cơ sở dữ liệu và tương tác người – máy tính là nội dung cốt lõi trong giáo trình khoa học dữ liệu.

Vào những năm 1990, Jim Gray – người từng nhận giải Turing – đã dự đoán khoa học dữ liệu sẽ trở thành "mô hình thứ tư" của khoa học (sau mô hình thực nghiệm, lý thuyết và tính toán), trong đó khoa học dựa trên dữ liệu sẽ là xu hướng tất yếu. Ông khẳng định rằng "mọi thứ trong khoa học đều đang thay đổi do tác động của công nghệ thông tin" và sự xuất hiện của dữ liệu khổng lồ. Do đó, không có gì ngạc nhiên khi khoa học dữ liệu, dữ liệu lớn và khai phá dữ liệu có mối quan hệ chặt chẽ với nhau và đại diện cho một xu hướng tất yếu trong sự phát triển của khoa học và công nghệ.

1.5.5 Khai phá dữ liệu và các lĩnh vực khác

Bên cạnh thống kê, học máy và công nghệ cơ sở dữ liệu, khai phá dữ liệu còn có mối quan hệ chặt chẽ với nhiều lĩnh vực khác.

Phần lớn dữ liệu trong thực tế có dạng phi cấu trúc, chẳng hạn như văn bản ngôn ngữ tự nhiên, hình ảnh hoặc dữ liệu âm thanh – video. Do đó, các lĩnh vực như xử lý ngôn ngữ tự nhiên, thị giác máy tính, nhận dạng mẫu, xử lý tín hiệu âm thanh – video, và phục hồi thông tin đóng vai trò quan trọng trong việc xử lý các loại dữ liệu này. Thực tế, để xử lý bất kỳ loại dữ liệu đặc thù nào, cần tích hợp nhiều lĩnh vực kiến thức vào thiết kế thuật toán khai phá dữ liệu. Chẳng hạn, khai phá dữ liệu y sinh yêu cầu sự kết hợp giữa các lĩnh vực khoa học sinh học, y học và tin sinh học. Khai phá dữ liệu không gian địa lý đòi hỏi kỹ thuật và kiến thức từ địa lý và khoa học dữ liệu không gian địa lý. Phát hiện lỗi phần mềm trong các hệ thống phần mềm lớn cần sự kết hợp giữa kỹ thuật phần mềm và khai phá dữ liệu. Khai phá dữ liệu mạng xã hội và truyền thông xã hội đòi hỏi kiến thức từ khoa học xã hội và khoa học mạng. Những ví dụ này còn có thể mở rộng mãi vì khai phá dữ liệu đã và đang xâm nhập vào hầu hết các lĩnh vực ứng dụng.

Một trong những thách thức lớn trong khai phá dữ liệu là đảm bảo hiệu suất và khả năng mở rộng, do chúng ta thường phải xử lý một lượng dữ liệu khổng lồ trong điều kiện giới hạn về thời gian và tài nguyên. Khai phá dữ liệu có liên quan chặt chẽ đến thiết kế thuật toán hiệu quả, chẳng hạn như các thuật toán có độ phức tạp thấp, thuật toán khai phá dữ liệu gia tăng (incremental), và thuật toán khai phá dữ liệu trên luồng dữ liệu. Ngoài ra, khai phá dữ liệu thường tận dụng sức mạnh của tính toán hiệu năng cao, tính toán song song và tính toán phân tán, cùng với các nền tảng phần cứng tiên tiến và môi trường điện toán đám mây hoặc cụm máy tính.

Khai phá dữ liệu cũng gắn bó chặt chẽ với lĩnh vực tương tác người – máy. Người dùng cần tương tác với hệ thống khai phá dữ liệu một cách hiệu quả, chẳng hạn như yêu cầu hệ thống khai phá dữ liệu gì, cách tích hợp kiến thức nền, cách thực hiện khai phá, và cách trình bày kết quả khai phá sao cho dễ hiểu (ví dụ: bằng cách diễn giải và trực quan hóa) cũng như dễ tương tác (ví dụ: bằng giao diện đồ họa thân thiện hoặc khai phá dữ liệu tương tác).

Thực tế hiện nay không chỉ có nhiều hệ thống khai phá dữ liệu tương tác mà còn có rất nhiều chức năng khai phá dữ liệu được tích hợp ẩn trong các chương trình ứng dụng khác nhau. Không thực tế nếu mong đợi mọi người trong xã hội đều hiểu và thành thạo kỹ thuật khai phá dữ liệu, cũng như không thể để các tổ chức công khai bộ dữ liệu khổng lồ của họ. Do đó, nhiều hệ thống đã tích hợp các chức năng khai phá dữ liệu để người dùng có thể thực hiện khai phá dữ liệu hoặc sử dụng kết quả khai phá chỉ bằng một cú nhấp chuột. Ví dụ, các công cụ tìm kiếm thông minh và các nền tảng bán lẻ trực tuyến thực hiện **khai phá dữ liệu ẩn** bằng cách thu thập dữ liệu của họ và lịch sử tìm kiếm hoặc mua sắm của người dùng, từ

đó cải thiện hiệu suất, chức năng và trải nghiệm người dùng. Khi người dùng mua sắm trực tuyến, có thể ngạc nhiên khi nhận được các gợi ý thông minh. Đó rất có thể là kết quả của các kỹ thuật khai phá dữ liệu ẩn này.

1.6 Khai phá dữ liệu và ứng dụng

Nơi nào có dữ liệu, nơi đó có các ứng dụng khai phá dữ liệu.

Là một lĩnh vực có hướng ứng dụng rất cao, khai phá dữ liệu đã gặt hái được nhiều thành công lớn trong nhiều ứng dụng khác nhau. Thật không thể liệt kê hết tất cả các ứng dụng mà khai phá dữ liệu đóng vai trò then chốt. Việc trình bày khai phá dữ liệu trong các lĩnh vực ứng dụng đòi hỏi nhiều tri thức chuyên sâu, chẳng hạn như trong tin sinh học hay kỹ thuật phần mềm, vượt ra ngoài phạm vi của cuốn sách này. Để minh họa tầm quan trọng của các ứng dụng khai phá dữ liệu, chúng tôi sẽ thảo luận một cách ngắn gọn về một vài ví dụ ứng dụng nổi bật và thành công của khai phá dữ liệu, bao gồm: *trí tuệ kinh doanh*; các *công cụ tìm kiếm*; *truyền thông xã hội* và các *mạng xã hội*; cũng như các ứng dụng trong *sinh học*, *y học* và *chăm sóc sức khỏe*.

1.6.1 Trí tuệ kinh doanh

Việc doanh nghiệp nắm bắt được bối cảnh thương mại của tổ chức, chẳng hạn như khách hàng, thị trường, nguồn cung và tài nguyên, cũng như đối thủ cạnh tranh, là vô cùng quan trọng. Các công nghệ **trí tuệ kinh doanh** (business intelligence, BI) cung cấp cái nhìn tổng quan về hoạt động kinh doanh qua ba góc độ: lịch sử, hiện tại và dự báo trong tương lai. Ví dụ điển hình bao gồm báo cáo, xử lý phân tích trực tuyến, quản lý hiệu suất kinh doanh, thu thập thông tin cạnh tranh, so sánh với chuẩn mực (benchmarking) và phân tích dự báo.

"*Tầm quan trọng của khai phá dữ liệu trong trí tuệ kinh doanh là gì?*" Nếu thiếu khai phá dữ liệu, nhiều doanh nghiệp có thể sẽ không thể thực hiện các phân tích thị trường hiệu quả, so sánh phản hồi khách hàng về các sản phẩm tương tự, phát hiện ra điểm mạnh và điểm yếu của đối thủ cạnh tranh, giữ chân được những khách hàng có giá trị cao và đưa ra các quyết định kinh doanh thông minh.

Rõ ràng, khai phá dữ liệu là cốt lõi của trí tuệ kinh doanh. Các công cụ xử lý phân tích trực tuyến trong BI dựa vào kho dữ liệu và khai phá dữ liệu nhiều chiều. Các kỹ thuật phân loại và dự đoán là trung tâm của phân tích dự báo trong BI,

được ứng dụng rộng rãi trong việc phân tích thị trường, nguồn cung và doanh số bán hàng. Hơn nữa, phân cụm đóng vai trò quan trọng trong quản lý mối quan hệ khách hàng, khi nó nhóm các khách hàng dựa trên các đặc trưng tương đồng của họ. Nhờ các kỹ thuật tóm tắt dữ liệu nhiều chiều, chúng ta có thể hiểu rõ hơn các đặc điểm của từng nhóm khách hàng và phát triển các chương trình thưởng, khuyến mãi được cá nhân hoá phù hợp.

1.6.2 Công cụ tìm kiếm trên Web

Một **công cụ tìm kiếm trên Web** là một máy chủ máy tính chuyên dụng dùng để tìm kiếm thông tin trên Internet. Kết quả của một truy vấn từ người dùng thường được trả về dưới dạng một danh sách (đôi khi gọi là *hits*). Những "hits" này có thể bao gồm các trang web, hình ảnh và các loại tệp tin khác. Một số công cụ tìm kiếm còn tìm kiếm và trả về dữ liệu từ các cơ sở dữ liệu công cộng hoặc thư mục mở. Sự khác biệt giữa công cụ tìm kiếm và **thư mục Web** nằm ở chỗ thư mục Web được duy trì bởi biên tập viên con người, trong khi các công cụ tìm kiếm hoạt động theo thuật toán hoặc sự kết hợp giữa thuật toán và đầu vào từ phía con người.

Các công cụ tìm kiếm đặt ra nhiều thách thức lớn đối với khai phá dữ liệu. Trước hết, các công cụ tìm kiếm phải xử lý một lượng dữ liệu rất lớn và không ngừng tăng. Thông thường, dữ liệu như vậy không thể xử lý chỉ bằng một hoặc vài máy tính. Thay vào đó, các công cụ tìm kiếm thường sử dụng các *cụm máy tính* (computer clouds) gồm hàng nghìn hoặc thậm chí hàng trăm nghìn máy tính làm việc cùng nhau để khai thác khối lượng dữ liệu khổng lồ. Việc mở rộng các phương pháp khai phá dữ liệu trên các cụm máy tính và các tập dữ liệu phân tán lớn là một lĩnh vực đang được nghiên cứu và phát triển tích cực.

Thứ hai, công cụ tìm kiếm thường phải xử lý dữ liệu trực tuyến. Một công cụ tìm kiếm có thể xây dựng một mô hình dựa trên dữ liệu khổng lồ trong quá trình làm việc ngoại tuyến. Ví dụ, nó có thể xây dựng một bộ phân loại truy vấn (query classifier) để gán một truy vấn tìm kiếm vào các danh mục đã định sẵn dựa trên chủ đề của truy vấn (ví dụ: truy vấn "apple" có ý nghĩa là tìm thông tin về loại trái cây hay thương hiệu máy tính). Ngay cả khi mô hình được xây dựng ngoại tuyến, việc điều chỉnh mô hình trực tuyến phải đủ nhanh để trả lời truy vấn của người dùng theo thời gian thực.

Một thách thức khác là duy trì và cập nhật mô hình một cách gia tăng trên các luồng dữ liệu liên tục tăng nhanh. Ví dụ, bộ phân loại truy vấn cần được duy trì liên tục vì các truy vấn mới luôn xuất hiện, và các danh mục định sẵn cũng như phân

phối dữ liệu có thể thay đổi. Hầu hết các phương pháp huấn luyện mô hình hiện nay đều là ngoại tuyến và tĩnh, do đó không phù hợp với kịch bản này.

Thứ ba, công cụ tìm kiếm thường phải đối mặt với các truy vấn chỉ được hỏi rất ít lần. Giả sử một công cụ tìm kiếm muốn cung cấp các gợi ý truy vấn *dựa trên ngữ cảnh*. Khi người dùng đặt truy vấn, công cụ tìm kiếm sẽ cố gắng suy ra ngữ cảnh của truy vấn dựa trên hồ sơ người dùng và lịch sử truy vấn của họ để trả về các kết quả tùy chỉnh trong một khoảng thời gian rất ngắn. Tuy nhiên, mặc dù tổng số truy vấn có thể rất lớn, nhiều truy vấn lại chỉ được hỏi một hoặc vài lần. Những dữ liệu có phân bố lệch nghiêm trọng như vậy đặt ra thách thức lớn cho nhiều phương pháp khai phá dữ liệu và học máy.

1.6.3 Mạng xã hội và truyền thông xã hội

Sự phổ biến của mạng xã hội và truyền thông xã hội đã làm thay đổi căn bản cuộc sống của chúng ta cũng như cách chúng ta trao đổi thông tin và giao tiếp xã hội ngày nay. Với khối lượng dữ liệu khổng lồ từ mạng xã hội và truyền thông xã hội, việc phân tích dữ liệu này để trích xuất các mẫu và xu hướng có thể hành động là vô cùng quan trọng.

Khai phá dữ liệu truyền thông xã hội là quá trình sàng lọc khối lượng lớn dữ liệu truyền thông xã hội (chẳng hạn như dữ liệu về cách người dùng sử dụng mạng xã hội, hành vi trực tuyến, mối quan hệ giữa các cá nhân, hành vi mua sắm trực tuyến, trao đổi nội dung, v.v.) nhằm nhận diện các mẫu và xu hướng. Những mẫu và xu hướng này đã được ứng dụng vào nhiều lĩnh vực, bao gồm phát hiện sự kiện xã hội, giám sát và theo dõi sức khỏe cộng đồng, phân tích cảm xúc trên mạng xã hội, hệ thống gợi ý trong mạng xã hội, truy xuất nguồn gốc thông tin, phân tích độ tin cậy của dữ liệu trên mạng xã hội, và phát hiện kẻ gửi thư rác trên mạng xã hội.

Khai phá dữ liệu mạng xã hội là nghiên cứu cấu trúc của mạng xã hội và thông tin liên quan đến các mạng này thông qua việc sử dụng lý thuyết đồ thị và các phương pháp khai phá dữ liệu. Cấu trúc của mạng xã hội được đặc trưng bởi các nút (các cá nhân, con người hoặc thực thể trong mạng) và các cạnh, hay các liên kết (quan hệ hoặc tương tác) kết nối giữa chúng. Một số ví dụ về cấu trúc xã hội thường được mô tả thông qua phân tích mạng xã hội bao gồm: mạng truyền thông xã hội, sự lan truyền ảnh chế (meme), mạng quan hệ bạn bè và người quen, đồ thị hợp tác, quan hệ huyết thống, sự lây truyền bệnh, và quan hệ tình dục. Những mạng này thường được trực quan hóa thông qua đồ thị xã hội (sociogram), trong đó các nút được biểu diễn dưới dạng điểm và các mối liên kết được biểu diễn dưới

dạng đường nối.

Khai phá dữ liệu mạng xã hội đã được sử dụng để phát hiện cộng đồng ẩn, nghiên cứu sự phát triển và động lực của mạng xã hội, tính toán các chỉ số mạng (chẳng hạn như tính trung tâm, tính bắc cầu, tính tương hỗi hay đối ứng, cân bằng, trạng thái và mức độ tương đồng), phân tích sự lan truyền thông tin trên các trang mạng xã hội, đo lường và mô hình hóa ảnh hưởng của nút / cấu trúc con và hiện tượng đồng nhất (homophily), cũng như thực hiện phân tích mạng xã hội dựa trên vị trí địa lý.

Khai phá dữ liệu truyền thông xã hội và mạng xã hội là những ứng dụng quan trọng của khai phá dữ liệu.

1.6.4 Sinh học, y học và chăm sóc sức khỏe

Sinh học, y học và chăm sóc sức khỏe cũng đang tạo ra một lượng dữ liệu khổng lồ với tốc độ tăng theo cấp số nhân. Dữ liệu y sinh có nhiều dạng khác nhau, từ dữ liệu "omics" (genomics – nghiên cứu về bộ gen, proteomics – nghiện cứu về protein, metabolomics – nghiên cứu về chuyển hóa, v.v.) đến hình ảnh y khoa, dữ liệu y tế di động và hồ sơ sức khỏe điện tử. Nhờ vào các phương pháp thu thập dữ liệu kỹ thuật số ngày càng hiệu quả, các nhà khoa học y sinh và bác sĩ lâm sàng hiện nay đang phải đối mặt với tập dữ liệu ngày càng lớn, đồng thời tìm cách sáng tạo để xử lý và phân tích khối lượng dữ liệu khổng lồ này. Trên thực tế, những tập dữ liệu từng được coi là lớn trước đây giờ đây lại trở nên nhỏ bé khi lượng dữ liệu mà một nhà nghiên cứu có thể thu thập chỉ trong một ngày có thể vượt xa tổng lượng dữ liệu mà họ từng thu thập trong suốt sự nghiệp của mình cách đây một thập kỷ. Sự bùng nổ dữ liệu y sinh này đòi hỏi những cách tư duy mới về cách quản lý và phân tích dữ liệu nhằm nâng cao hiểu biết khoa học và cải thiện chăm sóc sức khỏe.

Khai phá dữ liệu y sinh bao gồm nhiều bài toán khai phá dữ liệu đầy thách thức, bao gồm khai phá dữ liệu trình tự gen và protein quy mô lớn, khai phá mẫu đồ thị con phổ biến để phân loại dữ liệu sinh học, khai phá mạng điều hòa sinh học, mô tả và dự đoán tương tác protein – protein, phân loại và phân tích dự đoán trên ảnh y khoa, khai phá văn bản sinh học, xây dựng mạng thông tin sinh học từ dữ liệu văn bản y sinh, khai phá hồ sơ sức khỏe điện tử, khai phá mạng y sinh. Việc áp dụng khai phá dữ liệu vào lĩnh vực y sinh không chỉ giúp cải thiện chất lượng chẩn đoán và điều trị mà còn đóng góp quan trọng vào nghiên cứu khoa học và phát triển các phương pháp điều trị tiên tiến.

1.7 Khai phá dữ liệu và xã hội

Khi khai phá dữ liệu ngày càng thâm nhập vào đời sống hằng ngày, việc nghiên cứu tác động của nó đối với xã hội trở nên quan trọng. Chúng ta có thể sử dụng công nghệ khai phá dữ liệu để mang lại lợi ích cho xã hội như thế nào? Làm sao để ngăn chặn việc lạm dụng công nghệ này? Những vấn đề như tiết lộ hoặc sử dụng dữ liệu không đúng cách, cũng như nguy cơ vi phạm quyền riêng tư cá nhân và bảo vệ dữ liệu là những mối quan tâm cần được giải quyết.

Khai phá dữ liệu có thể hỗ trợ khám phá khoa học, quản lý doanh nghiệp, phục hồi kinh tế và bảo vệ an ninh (ví dụ: phát hiện kẻ xâm nhập và các cuộc tấn công mạng trong thời gian thực). Tuy nhiên, nó cũng có thể vô tình tiết lộ thông tin mật của doanh nghiệp hoặc chính phủ, cũng như thông tin cá nhân của mọi người. Do đó, các nghiên cứu về bảo mật dữ liệu trong khai phá dữ liệu và phương pháp xuất bản dữ liệu cũng như khai phá dữ liệu bảo vệ quyền riêng tư là những chủ đề nghiên cứu quan trọng và đang được quan tâm. Triết lý cốt lõi là phải đảm bảo tính nhạy cảm của dữ liệu, duy trì bảo mật dữ liệu và quyền riêng tư của con người trong khi vẫn thực hiện khai phá dữ liệu thành công.

Những vấn đề này, cùng với nhiều vấn đề khác liên quan đến nghiên cứu, phát triển và ứng dụng khai phá dữ liệu, sẽ được thảo luận xuyên suốt cuốn sách này.

1.8 Tóm tắt

- *Nhu cầu là nguồn gốc của phát minh.* Với sự gia tăng mạnh mẽ của dữ liệu trong mọi ứng dụng, khai phá dữ liệu đáp ứng nhu cầu cấp thiết về phân tích dữ liệu hiệu quả, có khả năng mở rộng và linh hoạt trong xã hội. Khai phá dữ liệu có thể được xem là một sự tiến hóa tự nhiên của công nghệ thông tin, đồng thời là sự hội tụ của nhiều ngành liên quan và lĩnh vực ứng dụng khác nhau.

- **Khai phá dữ liệu** là quá trình khám phá các mẫu và tri thức thú vị từ một lượng dữ liệu khổng lồ. Như một *quy trình khám phá tri thức*, nó thường bao gồm các bước: làm sạch dữ liệu, tích hợp dữ liệu, chọn lọc dữ liệu, biến đổi dữ liệu, khám phá mẫu và mô hình, đánh giá mẫu hoặc mô hình, và trình bày tri thức.

- Một mẫu hoặc mô hình được coi là *thú vị* nếu nó có hiệu lực trên dữ liệu kiểm tra ở một mức độ nhất định, mới lạ, có tiềm năng hữu ích (ví dụ: có thể

áp dụng hoặc xác nhận một giả thuyết mà người dùng quan tâm), và dễ hiểu đối với con người. Các mẫu thú vị đại diện cho tri thức. Các thước đo **mức độ thú vị của mẫu**, có thể là *khách quan* hoặc *chủ quan*, giúp định hướng quá trình khám phá.

- Khai phá dữ liệu có thể được thực hiện trên bất kỳ loại **dữ liệu** nào miễn là dữ liệu đó có ý nghĩa đối với một ứng dụng cụ thể, bao gồm dữ liệu có cấu trúc (ví dụ: cơ sở dữ liệu quan hệ, dữ liệu giao dịch) và dữ liệu phi cấu trúc (ví dụ: văn bản và dữ liệu đa phương tiện), cũng như dữ liệu từ các ứng dụng khác nhau. Dữ liệu cũng có thể được phân loại thành dữ liệu lưu trữ và dữ liệu luồng, trong đó dữ liệu luồng có thể yêu cầu các thuật toán khai phá dữ liệu đặc biệt.

- Các **chức năng khai phá dữ liệu** được sử dụng để xác định các loại mẫu hoặc **tri thức** cần tìm trong các tác vụ khai phá dữ liệu. Các chức năng này bao gồm đặc trưng hóa và phân biệt; khai phá các mẫu thường xuyên, các luật kết hợp và mối tương quan; phân loại và hồi quy; học sâu; phân cụm; và phát hiện ngoại lệ. Khi xuất hiện các loại dữ liệu mới, các ứng dụng mới và nhu cầu phân tích mới, chắc chắn sẽ có nhiều tác vụ khai phá dữ liệu mới được phát triển trong tương lai.

- Khai phá dữ liệu là sự hội tụ của nhiều ngành nhưng vẫn có trọng tâm nghiên cứu riêng biệt, hướng đến nhiều ứng dụng tiên tiến. Chúng ta đã nghiên cứu mối quan hệ mật thiết giữa khai phá dữ liệu với thống kê, học máy, công nghệ cơ sở dữ liệu và nhiều lĩnh vực khác.

- Khai phá dữ liệu có nhiều **ứng dụng** thành công, chẳng hạn như trong trí tuệ kinh doanh, tìm kiếm web, tin sinh học, tin học y tế, tài chính, thư viện số và chính phủ số.

- Khai phá dữ liệu đã và đang có tác động mạnh mẽ đến xã hội, và việc nghiên cứu những tác động này, chẳng hạn như làm thế nào để đảm bảo hiệu quả khai phá dữ liệu đồng thời bảo vệ quyền riêng tư và bảo mật dữ liệu, đã trở thành một vấn đề nghiên cứu quan trọng.

1.9 Bài tập

1. *Khai phá dữ liệu* là gì? Trong câu trả lời của bạn, hãy đề cập đến các nội dung sau:

a Liệu khai phá dữ liệu có chỉ đơn giản là một sự chuyển đổi hoặc ứng dụng của công nghệ được phát triển từ các lĩnh vực *cơ sở dữ liệu*, *thống kê*, *học máy* và *nhận dạng mẫu* không?

b Một số người tin rằng khai phá dữ liệu là kết quả tất yếu của sự phát triển của công nghệ thông tin. Nếu bạn là một nhà nghiên cứu cơ sở dữ liệu, hãy chứng minh rằng khai phá dữ liệu là một sự tiến hóa tự nhiên của công nghệ cơ sở dữ liệu. Nếu bạn là một nhà nghiên cứu học máy hoặc một nhà thống kê thì sao?

c Mô tả các bước liên quan đến khai phá dữ liệu khi được xem như một quá trình khám phá tri thức.

2. Định nghĩa từng *chức năng của khai phá dữ liệu*: phân tích luật kết hợp và tương quan, phân loại, hồi quy, phân cụm, học sâu và phân tích ngoại lệ. Đưa ra ví dụ cho từng chức năng khai phá dữ liệu, sử dụng một cơ sở dữ liệu thực tế mà bạn quen thuộc.

3. Trình bày một ví dụ mà khai phá dữ liệu đóng vai trò quan trọng trong sự thành công của một doanh nghiệp. Doanh nghiệp này cần những *chức năng khai phá dữ liệu* nào (ví dụ: nghĩ về các loại mẫu có thể được khai phá)? Liệu những mẫu này có thể được tạo ra theo cách khác bằng cách xử lý truy vấn dữ liệu hoặc phân tích thống kê đơn giản không?

4. Giải thích sự khác biệt và tương đồng giữa các khái niệm sau: phân tích tương quan và phân loại, phân loại và phân cụm, phân loại và hồi quy.

5. Dựa trên quan sát của bạn, hãy mô tả một loại tri thức khác có thể cần được khám phá bằng các phương pháp khai phá dữ liệu nhưng chưa được liệt kê trong chương này.** Loại tri thức này có yêu cầu một phương pháp khai phá hoàn toàn khác so với các phương pháp đã được đề cập không?

6. Điểm *ngoại lệ* thường bị loại bỏ như là nhiễu. Tuy nhiên, rác của người này có thể là kho báu của người khác. Ví dụ, các giao dịch ngoại lệ trong thẻ tín dụng có thể giúp phát hiện gian lận sử dụng thẻ tín dụng. Sử dụng phát hiện gian lận làm ví dụ, đề xuất hai phương pháp có thể được sử dụng để phát hiện điểm ngoại lệ và thảo luận phương pháp nào đáng tin cậy hơn.

7. Những thách thức chính trong việc khai phá một lượng dữ liệu khổng lồ (ví dụ: hàng tỷ bộ dữ liệu) so với việc khai phá một lượng dữ liệu nhỏ (ví dụ: vài trăm bộ dữ liệu) là gì?

8. Phác thảo các thách thức nghiên cứu chính trong khai phá dữ liệu đối với một lĩnh vực ứng dụng cụ thể, chẳng hạn như phân tích dữ liệu luồng / sensor, phân tích dữ liệu không gian – thời gian hoặc tin sinh học.

1.10 Tài liệu tham khảo

Cuốn sách *Knowledge Discovery in Databases*, do Piatetsky – Shapiro và Frawley [PSF91] biên tập, là một trong những tuyển tập nghiên cứu đầu tiên về khám phá tri thức từ dữ liệu. Cuốn *Advances in Knowledge Discovery and Data Mining*, biên tập bởi Fayyad, Piatetsky – Shapiro, Smyth, và Uthurusamy [FPSSe96], là một tuyển tập nghiên cứu sớm khác về khám phá tri thức và khai phá dữ liệu. Kể từ đó, nhiều sách giáo khoa và sách nghiên cứu về khai phá dữ liệu đã được xuất bản. Một số cuốn phổ biến bao gồm: *Data Mining: Practical Machine Learning Tools and Techniques* (tái bản lần thứ 4) của Witten, Frank, Hall và Pal [WFHP16]; *Data Mining: Concepts and Techniques* (tái bản lần thứ 3) của Han, Kamber và Pei [HKP11]; *Introduction to Data Mining* (tái bản lần 2) của Tan, Steinbach, Karpatne và Kumar [TSKK18]; *Data Mining: The Textbook* của Aggarwal [Agg15b]; *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (tái bản lần 2nd) của Zaki và Meira [ZJ20]; *Mining of Massive Datasets* (tái bản lần 3) của Leskovec, Rajaraman và Ullman [ZJ20]; *The Elements of Statistical Learning* (tái bản lần 2) của Hastie, Tibshirani và Friedman [HTF09]; *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (tái bản lần 3) của Linoff và Berry [LB11]; *Principles of Data Mining (Adaptive Computation and Machine Learning)* của Hand, Mannila và Smyth [HMS01]; *Mining the Web: Discovering Knowledge from Hypertext Data* của Chakrabarti [Cha03]; *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* của Liu [Liu06]; *Data Mining: Multimedia, Soft Computing, and Bioinformatics* của Mitra và Acharya [MA03].

Ngoài ra, có rất nhiều sách tổng hợp các bài báo hoặc chương chuyên sâu về các khía cạnh cụ thể của khám phá tri thức, chẳng hạn như phân cụm, phát hiện điểm ngoại lệ, phân loại, khai phá luật kết hợp, và khai phá dữ liệu từ các loại dữ liệu đặc thù như văn bản, đa phương tiện, dữ liệu quan hệ, dữ liệu không gian, dữ liệu mạng xã hội và dữ liệu phương tiện truyền thông xã hội. Tuy nhiên, danh sách này đã trở nên rất dài theo thời gian, và chúng tôi không liệt kê tất cả. Có nhiều tài liệu hướng dẫn về khai phá dữ liệu trong các hội nghị lớn về khai phá dữ liệu, cơ sở dữ liệu, học máy, thống kê và công nghệ Web.

KDNuggets là một bản tin điện tử thường xuyên chứa các thông tin liên quan đến khám phá tri thức và khai phá dữ liệu, được điều hành bởi Piatetsky – Shapiro từ năm 1991. Trang web *KDNuggets* (https://www.kdnuggets.com) cũng cung cấp một bộ sưu tập thông tin tốt về KDD.

Cộng đồng khai phá dữ liệu đã tổ chức hội nghị quốc tế đầu tiên về khám phá tri thức và khai phá dữ liệu vào năm 1995. Hội nghị này phát triển từ bốn hội thảo quốc tế về khám phá tri thức trong cơ sở dữ liệu, được tổ chức từ năm 1989 đến 1994. Nhóm đặc biệt ACM – SIGKDD (Special Interest Group on Knowledge Discovery in Databases) được thành lập dưới sự bảo trợ của ACM vào năm 1998 và từ năm 1999 đã tổ chức các hội nghị quốc tế về khám phá tri thức và khai phá dữ liệu. Hiệp hội Khoa học Máy tính IEEE đã tổ chức hội nghị khai phá dữ liệu hàng năm của mình, “International Conference on Data Mining (ICDM)”, từ năm 2001. Hiệp hội Toán học Công nghiệp và Ứng dụng (Society on Industrial and Applied Mathematics, SIAM) đã tổ chức hội nghị khai phá dữ liệu hàng năm, “SIAM Data Mining Conference (SDM)”, từ năm 2002. Tạp chí chuyên biệt *Data Mining and Knowledge Discovery*, do Springer xuất bản, đã có từ năm 1997. Tạp chí ACM, *ACM Transactions on Knowledge Discovery from Data*, ra mắt số đầu tiên vào năm 2007.

ACM – SIGKDD cũng xuất bản bản tin hai năm một lần, *SIGKDD Explorations*. Có một số hội nghị quốc tế hoặc khu vực khác về khai phá dữ liệu, chẳng hạn như: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), International Conference on Web Search and Data Mining (WSDM).

Nghiên cứu về khai phá dữ liệu cũng được công bố rộng rãi trong nhiều sách giáo khoa, sách nghiên cứu, hội nghị và tạp chí về khai phá dữ liệu, cơ sở dữ liệu, thống kê, học máy và trực quan hóa dữ liệu.

Chương 2

Dữ liệu, phép đo và tiền xử lý dữ liệu

2.1 Kiểu dữ liệu

2.1.1 Thuộc tính danh định

2.1.2 Thuộc tính nhị phân

2.1.3 Thuộc tính thứ tự

2.1.4 Thuộc tính số

2.1.5 Thuộc tính rời rạc và liên tục

2.2 Thống kê dữ liệu

2.2.1 Đo lường xu hướng trung tâm

2.2.2 Đo lường độ phân tán của dữ liệu

2.2.3 Phân tích hiệp phương sai và tương quan

2.2.4 Biểu diễn đồ họa của thống kê cơ bản

2.3 Độ tương đồng và khoảng cách

2.3.1 Ma trận dữ liệu và ma trận độ không tương đồng

2.3.2 Thước đo gần gũi cho thuộc tính danh định

2.3.3 Thước đo gần gũi cho thuộc tính nhị phân

2.3.4 Độ không tương đồng của dữ liệu số: khoảng cách Minkowski

2.3.5 Thước đo gần gũi cho thuộc tính thứ tự

2.3.6 Độ không tương đồng cho thuộc tính kiểu hỗn hợp

2.3.7 Độ tương đồng cosine

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

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

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

2.4.1 Thước đo chất lượng dữ liệu

2.4.2 Làm sạch dữ liệu

2.4.3 Tích hợp dữ liệu

2.5 Biến đổi dữ liệu

2.5.1 Chuẩn hóa

2.5.2 Rời rạc hóa

2.5.3 Nén dữ liệu

2.5.4 Lấy mẫu

2.6 Giảm chiều dữ liệu

2.6.1 Phân tích thành phần chính

2.6.2 Lựa chọn tập con thuộc tính

2.6.3 Các phương pháp giảm chiều phi tuyến

2.7 Tóm tắt

2.8 Bài tập

2.9 Tài liệu tham khảo

Tài liệu tham khảo

1. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., và Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* Tập 457, trang 1012–1014 (2009).
2. Han, J., Pei, J., và Tong, H. *Data Mining: Concepts and Techniques* In lần thứ 4. 786 trang (Morgan Kaufmann, 2022).