**3.4.Giảm dữ liệu**

Hãy tưởng tượng rằng bạn đã chọn dữ liệu từ kho dữ liệu AllElectronics để phân tích.

Bộ dữ liệu có thể sẽ rất lớn! Phân tích và khai thác dữ liệu phức tạp trên một lượng lớn

dữ liệu có thể mất nhiều thời gian, khiến việc phân tích như vậy không thực tế hoặc không khả thi.

Các kỹ thuật giảm thiểu dữ liệu có thể được áp dụng để có được sự đại diện giảm thiểu của tập dữ liệu có dung lượng nhỏ hơn nhiều nhưng vẫn duy trì chặt chẽ tính toàn vẹn của bản gốc dữ liệu. Nghĩa là, việc khai thác trên tập dữ liệu giảm sẽ hiệu quả hơn nhưng vẫn tạo ra kết quả phân tích giống nhau (hoặc gần giống nhau). Trong phần này, trước tiên chúng tôi trình bày tổng quan về của các chiến lược giảm thiểu dữ liệu, tiếp theo là xem xét kỹ hơn các kỹ thuật riêng lẻ.

**3.4.1. Tổng quan về chiến lược giảm dữ liệu**

Các chiến lược giảm dữ liệu bao gồm giảm kích thước, giảm số lượng và nén dữ liệu.

Giảm chiều là quá trình giảm số lượng các biến ngẫu nhiên hoặc các thuộc tính đang được xem xét. Các phương pháp giảm kích thước bao gồm: Tiền xử lý dữ liệu biến đổi (Phần 3.4.2) và phân tích các thành phần chính (Phần 3.4.3), chuyển đổi hoặc chiếu dữ liệu gốc lên một không gian nhỏ hơn. Lựa chọn tập hợp con thuộc tính là một phương pháp giảm kích thước trong đó không liên quan, kém phù hợp hoặc dư thừa các thuộc tính hoặc thứ nguyên được phát hiện và loại bỏ (Phần 3.4.4).Kỹ thuật giảm số lượng thay thế khối lượng dữ liệu ban đầu bằng phương pháp thay thế,các dạng biểu diễn dữ liệu nhỏ hơn. Các kỹ thuật này có thể là tham số hoặc phi tham số. Đối với phương pháp tham số, một mô hình được sử dụng để ước tính dữ liệu, do đó thường chỉ có các tham số dữ liệu cần được lưu trữ, thay vì dữ liệu thực tế. (Các giá trị ngoại lai cũng có thể được lưu trữ.) Các mô hình hồi quy và log-tuyến tính (Phần 3.4.5) là các ví dụ. Các phương pháp phi tham số để lưu trữ các đại diện đã giảm của dữ liệu bao gồm biểu đồ (Phần 3.4.6), phân nhóm (Phần 3.4.7), lấy mẫu (Phần 3.4.8) và dữ liệu tập hợp khối (Mục 3.4.9). Trong quá trình nén dữ liệu, các phép biến đổi được áp dụng để thu được phần biểu diễn thu nhỏ hoặc "nén" của dữ liệu gốc. Nếu dữ liệu ban đầu có thể được tái tạo từ dữ liệu nén mà không bị mất thông tin, việc giảm dữ liệu được gọi là không mất mát. Thay vào đó, nếu chúng tôi chỉ có thể tạo lại một số liệu gần đúng của dữ liệu ban đầu, thì giảm dữ liệu được gọi là mất mát. Có một số thuật toán không mất dữ liệu để nén chuỗi; tuy nhiên, chúng thường chỉ cho phép thao tác dữ liệu hạn chế. Kích thước các kỹ thuật giảm thiểu và giảm số lượng cũng có thể được coi là các dạng dữ liệu nén. Có nhiều cách khác để tổ chức các phương pháp giảm thiểu dữ liệu. Thời gian tính toán dành cho việc giảm dữ liệu không được lớn hơn hoặc "xóa" đi thời gian được lưu bởi khai thác trên kích thước tập dữ liệu giảm.

**3.4.2 Biến đổi Wavelet**

Biến đổi wavelet rời rạc (DWT) là một kỹ thuật xử lý tín hiệu tuyến tính, khi được áp dụng cho một vectơ dữ liệu X, biến đổi nó thành một vectơ khác về số, X0, của hệ số wavelet. Hai vectơ có cùng độ dài. Khi áp dụng kỹ thuật này để giảm dữ liệu, chúng tôi coi mỗi bộ giá trị như một vectơ dữ liệu n chiều là X =(x1, x2, …, xn ), mô tả n phép đo được thực hiện trên bộ từ n cơ sở dữ liệu thuộc tính. “Làm thế nào kỹ thuật này có thể hữu ích cho việc giảm dữ liệu nếu dữ liệu được chuyển đổi wavelet là có cùng độ dài với dữ liệu gốc? ” Tính hữu ích nằm ở thực tế là wavelet dữ liệu đã chuyển đổi có thể bị cắt bớt. Một xấp xỉ được nén của dữ liệu có thể là được giữ lại bằng cách chỉ lưu trữ một phần nhỏ của hệ số mạnh nhất trong số các hệ số wavelet.

Ví dụ: tất cả các hệ số wavelet lớn hơn một số ngưỡng do người dùng chỉ định có thể giữ lại. Tất cả các hệ số khác được đặt thành 0. Do đó, biểu diễn dữ liệu kết quả là rất thưa thớt, do đó các hoạt động có thể tận dụng sự thưa thớt dữ liệu được tính toán rất nhanh nếu được thực hiện trong không gian wavelet. Kỹ thuật này cũng hoạt động để loại bỏ nhiễu mà không làm mượt các tính năng chính của dữ liệu, giúp cho việc làm sạch dữ liệu cũng hiệu quả. Với một tập hợp các hệ số, giá trị gần đúng của dữ liệu ban đầu có thể là được xây dựng bằng cách áp dụng nghịch đảo của DWT được sử dụng.

DWT có liên quan chặt chẽ đến phép biến đổi Fourier rời rạc (DFT), một kỹ thuật xử lý tín hiệu liên quan đến sin và cosin. Nói chung, DWT đạt được tốt hơn nén mất mát. Nghĩa là, nếu cùng một số hệ số được giữ lại cho một DWT và DFT của một vectơ dữ liệu nhất định, phiên bản DWT sẽ cung cấp giá trị gần đúng chính xác hơn của dữ liệu gốc. Do đó, để có giá trị gần đúng tương đương, DWT yêu cầu ít hơn không gian hơn DFT. Không giống như DFT, các wavelet khá bản địa hóa trong không gian, góp phần để bảo tồn các chi tiết địa phương.

Chỉ có một DFT, nhưng có một số họ DWT. Hình 3.4 cho thấy một số họ wavelet. Các biến đổi wavelet phổ biến bao gồm Haar-2, Daubechies-4, và Daubechies-6. Quy trình chung để áp dụng biến đổi wavelet rời rạc sử dụng một thuật toán kim tự tháp phân cấp làm giảm một nửa dữ liệu ở mỗi lần lặp, dẫn đến tốc độ tính toán. Phương pháp như sau:

1. Độ dài, L, của vectơ dữ liệu đầu vào phải là lũy thừa nguyên của 2. Điều kiện này

có thể được đáp ứng bằng cách đệm véc tơ dữ liệu bằng các số không khi cần thiết (L ≥ n).

2. Mỗi phép biến đổi liên quan đến việc áp dụng hai hàm. Cách đầu tiên áp dụng một số làm mịn dữ liệu, chẳng hạn như tổng hoặc trung bình có trọng số. Thứ hai thực hiện một sự khác biệt có trọng số, hoạt động để làm nổi bật các tính năng chi tiết của dữ liệu.

3. Hai hàm được áp dụng cho các cặp điểm dữ liệu trong X, nghĩa là cho tất cả các cặp

số đo (x2i,x2i+1). Điều này dẫn đến hai tập dữ liệu có độ dài L = 2. Nói chung,

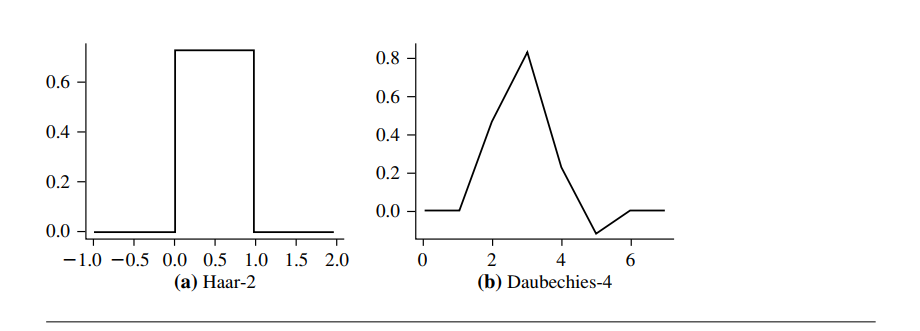
chúng đại diện cho phiên bản tần số thấp hoặc được làm mịn của dữ liệu đầu vào và nội dung tần số cao của nó, tương ứng.

4. Hai hàm được áp dụng đệ quy cho các tập dữ liệu thu được trong

vòng lặp, cho đến khi các tập dữ liệu kết quả thu được có độ dài 2.

5. Các giá trị đã chọn từ các tập dữ liệu thu được trong các lần lặp trước đó được chỉ định

hệ số wavelet của dữ liệu được biến đổi.



Tương tự, một phép nhân ma trận có thể được áp dụng cho dữ liệu đầu vào để thu được các hệ số wavelet, trong đó ma trận được sử dụng phụ thuộc vào DWT đã cho. Các ma trận phải là trực chuẩn, nghĩa là các cột là vectơ đơn vị và trực giao lẫn nhau, do đó nghịch đảo của ma trận chỉ là chuyển vị của nó. Mặc dù chúng tôi không có phòng để thảo luận về nó ở đây, thuộc tính này cho phép xây dựng lại dữ liệu từ các tập dữ liệu khác biệt trơn tru và mượt mà. Bằng cách tính toán ma trận được sử dụng thành một sản phẩm của một vài ma trận thưa thớt, kết quả là thuật toán “DWT nhanh” có độ phức tạp O(n) for một vectơ đầu vào có độ dài n. Các phép biến đổi Wavelet có thể được áp dụng cho dữ liệu đa chiều như một khối dữ liệu. Đây được thực hiện trước tiên bằng cách áp dụng phép biến đổi cho thứ nguyên đầu tiên, sau đó cho thứ hai, và như vậy trên. Độ phức tạp tính toán liên quan là tuyến tính đối với số lượng ô trong khối lập phương. Các phép biến đổi Wavelet cho kết quả tốt trên dữ liệu thưa thớt hoặc lệch và trên dữ liệu với các thuộc tính có thứ tự. Nén mất dữ liệu bằng wavelets được cho là tốt hơn JPEG

nén, tiêu chuẩn thương mại hiện tại. Biến đổi Wavelet có nhiều ứng dụng trong thế giới thực, bao gồm nén hình ảnh dấu vân tay, thị giác máy tính, phân tích dữ liệu chuỗi thời gian và làm sạch dữ liệu.

**3.4.3. Phép phân tích thành phần chính**

Giả sử rằng dữ liệu được giảm bớt bao gồm các bộ giá trị hoặc vectơ dữ liệu được mô tả bởi n thuộc tính hoặc thứ nguyên.

Phép phân tích thành phần chính (PCA) tìm kiếm k vectơ trực giao n-chiều có thể được sử dụng tốt nhất để biểu diễn dữ liệu, trong đó k ≤ n.

Do đó, dữ liệu ban đầu được chiếu vào một không gian nhỏ hơn nhiều, dẫn đến giảm kích thước.

Không giống như Trích chọn đặc trưng mà làm giảm kích thước tập hợp thuộc tính bằng cách giữ lại một tập hợp con của tập thuộc tính ban đầu, PCA “kết hợp” bản chất của các thuộc tính bằng cách tạo một tập hợp biến thay thế nhỏ hơn.

Dữ liệu ban đầu sau đó có thể được chiếu vào tập hợp nhỏ hơn này.

PCA thường tiết lộ các mối quan hệ mà trước đây không bị nghi ngờ và do đó cho phép các diễn giải mà thông thường không dẫn đến kết quả.

1. Dữ liệu đầu vào được chuẩn hóa để mỗi thuộc tính nằm trong cùng một phạm vi. Bước này giúp đảm bảo rằng các thuộc tính có miền lớn sẽ không lấn át các thuộc tính có miền nhỏ hơn.

2. PCA tính k vectơ trực chuẩn cung cấp cơ sở cho dữ liệu đầu vào chuẩn hóa. Đây là các vectơ đơn vị mà mỗi điểm theo phương vuông góc với các vectơ khác. Các vectơ này được gọi là các thành phần chính. Dữ liệu đầu vào là sự kết hợp tuyến tính của các thành phần chính.

3. Các thành phần chính được sắp xếp theo thứ tự giảm dần “mức độ quan trọng” hoặc độ mạnh. Các thành phần chính về cơ bản đóng vai trò như một tập hợp các trục mới cho dữ liệu, cung cấp thông tin quan trọng về phương sai. Nghĩa là, các trục được sắp xếp sao cho trục đầu tiên hiển thị phương sai nhiều nhất trong số dữ liệu, trục thứ hai hiển thị phương sai cao nhất tiếp theo, v.v. Ví dụ, Hình 3.5 cho thấy hai thành phần chính đầu tiên, Y1 và Y2, cho tập dữ liệu đã cho ban đầu được ánh xạ tới các trục X1 và X2. Thông tin này giúp xác định các nhóm hoặc mẫu trong dữ liệu.

4. Bởi vì các thành phần được sắp xếp theo thứ tự giảm dần về “mức độ quan trọng”, nên kích thước dữ liệu có thể được giảm xuống bằng cách loại bỏ các thành phần yếu hơn, tức là những thành phần có độ lớn biến thiên thấp. Sử dụng các thành phần chính mạnh nhất, sẽ có thể tạo lại một số liệu gần đúng tốt của dữ liệu ban đầu.

PCA có thể được áp dụng cho các thuộc tính có thứ tự và không có thứ tự, đồng thời có thể xử lý dữ liệu thưa thớt và dữ liệu lệch. Dữ liệu đa chiều của nhiều hơn hai chiều có thể được xử lý bằng cách giảm vấn đề xuống hai chiều. Các thành phần chính có thể được sử dụng làm đầu vào cho phân tích hồi quy nhiều lần và phân tích theo cụm. So với các phép biến đổi wavelet, PCA có xu hướng xử lý dữ liệu thưa thớt tốt hơn, trong khi các phép biến đổi wavelet phù hợp hơn với dữ liệu có kích thước cao

**3.4.4. Trích chọn đặc trưng**

(nói thêm)

Các tập dữ liệu để phân tích có thể chứa hàng trăm thuộc tính, nhiều thuộc tính có thể không phù hợp với nhiệm vụ khai thác hoặc dư thừa. Ví dụ: nếu nhiệm vụ là phân loại khách hàng dựa trên việc họ có khả năng mua một đĩa CD mới phổ biến tại AllElectronics hay không khi được thông báo về việc giảm giá, thì các thuộc tính như số điện thoại của khách hàng có thể không liên quan, không giống như các thuộc tính như tuổi hoặc gu âm nhạc. Mặc dù chuyên gia miền có thể chọn ra một số thuộc tính hữu ích, nhưng đây có thể là một nhiệm vụ khó khăn và tốn thời gian, đặc biệt là khi hành vi của dữ liệu không được biết rõ. Việc loại bỏ các thuộc tính có liên quan hoặc giữ lại các thuộc tính không liên quan có thể gây bất lợi, gây nhầm lẫn cho thuật toán khai thác được sử dụng. Điều này có thể dẫn đến việc phát hiện ra các mẫu có chất lượng kém. Ngoài ra, khối lượng bổ sung của các thuộc tính không liên quan hoặc dư thừa có thể làm chậm quá trình khai thác

(có thể làm power point từ phần này)

Trích chọn đặc trưng làm giảm kích thước tập dữ liệu bằng cách loại bỏ các thuộc tính (hoặc thứ nguyên) không liên quan hoặc dư thừa. Mục tiêu của Trích chọn đặc trưng là tìm một tập hợp tối thiểu các thuộc tính sao cho phân phối xác suất kết quả của các lớp dữ liệu càng gần với phân phối ban đầu càng tốt bằng cách sử dụng tất cả các thuộc tính. Khai thác trên một tập hợp các thuộc tính giảm có một lợi ích bổ sung:

+ Nó làm giảm số lượng các thuộc tính xuất hiện trong các mẫu đã phát hiện, giúp làm cho các mẫu dễ hiểu hơn.

Đối với n thuộc tính, có thể có hai mũ n tập con. Một tìm kiếm toàn diện cho tập hợp con tối ưu của các thuộc tính có thể rất tốn kém, đặc biệt là khi n và số lượng lớp dữ liệu tăng lên. Do đó, các phương pháp suy nghiệm (heuristic methods) khám phá không gian tìm kiếm thu gọn thường được sử dụng để lựa chọn tập con thuộc tính. Các phương thức này tham lam ở chỗ, trong khi tìm kiếm trong không gian thuộc tính, chúng luôn làm những gì được cho là lựa chọn tốt nhất tại thời điểm đó. Chiến lược của các phương thức này là đưa ra một lựa chọn tối ưu cục bộ với hy vọng rằng điều này sẽ dẫn đến một giải pháp tối ưu trên toàn cầu. Các phương pháp tham lam như vậy có hiệu quả trong thực tế và có thể ước tính một giải pháp gần tối ưu.

Các thuộc tính “tốt nhất” (và “tồi tệ nhất”) thường được xác định bằng cách sử dụng các thử nghiệm có ý nghĩa thống kê, giả định rằng các thuộc tính độc lập với nhau. Nhiều biện pháp đánh giá thuộc tính khác có thể được sử dụng chẳng hạn như thước đo thu thập thông tin được sử dụng trong việc xây dựng cây quyết định để phân loại.

Các phương pháp suy nghiệm cơ bản của việc lựa chọn tập hợp con thuộc tính bao gồm các kỹ thuật sau đây, một số trong số đó được minh họa trong Hình 3.6

1. Lựa trọn chuyển tiếp từng bước (Stepwise forward selection): Thủ tục bắt đầu với một tập hợp rỗng các thuộc tính là tập hợp đã rút gọn. Các thuộc tính gốc tốt nhất được xác định và thêm vào tập hợp đã rút gọn. Ở mỗi bước hoặc lần lặp tiếp theo, các thuộc tính gốc tốt nhất còn lại sẽ được thêm vào tập hợp.

2. Lược bỏ ngược từng bước (Stepwise backward elimination): Thủ tục bắt đầu với tập hợp đầy đủ các thuộc tính. Ở mỗi bước, nó loại bỏ thuộc tính xấu nhất còn lại trong tập hợp.

3. Kết hợp giữa “Lựa trọn chuyển tiếp” và “Lược bỏ ngược”: Tại mỗi bước, thủ tục chọn thuộc tính tốt nhất và loại bỏ thuộc tính xấu nhất trong số các thuộc tính còn lại.

4. Cảm ứng cây quyết định (Decision tree induction): Các thuật toán cây quyết định ban đầu được dùng để phân loại. Quy nạp cây quyết định xây dựng một sơ đồ giống như cấu trúc trong đó mỗi nút bên trong (không phải lá) biểu thị một bài kiểm tra trên một thuộc tính, mỗi nhánh tương ứng với một kết quả của bài kiểm tra và mỗi nút bên ngoài (lá) biểu thị một dự đoán lớp. Tại mỗi nút, thuật toán chọn thuộc tính "tốt nhất" để phân vùng dữ liệu thành các lớp riêng lẻ.

Khi quy nạp cây quyết định được sử dụng để lựa chọn tập hợp con thuộc tính, một cây được cấu trúc từ dữ liệu đã cho. Tất cả các thuộc tính không xuất hiện trong cây được coi là không liên quan. Tập hợp các thuộc tính xuất hiện trong cây tạo thành tập hợp con rút gọn của các thuộc tính.

Các tiêu chí dừng cho các phương pháp có thể khác nhau. Quy trình có thể sử dụng một ngưỡng trên thước đo được sử dụng để xác định thời điểm dừng quá trình lựa chọn thuộc tính.

Trong một số trường hợp, chúng tôi có thể muốn tạo các thuộc tính mới dựa trên các thuộc tính khác. Việc xây dựng thuộc tính như vậy có thể giúp cải thiện độ chính xác và sự hiểu biết về cấu trúc trong dữ liệu có chiều cao. Ví dụ: chúng ta có thể muốn thêm vùng thuộc tính dựa trên chiều cao và chiều rộng của thuộc tính. Bằng cách kết hợp các thuộc tính, cấu trúc thuộc tính có thể khám phá thông tin còn thiếu về mối quan hệ giữa các thuộc tính dữ liệu, từ đó có thể hữu ích cho việc khám phá kiến thức.

**3.4.5. Mô hình hồi quy tuyến tính và log-linear: Giảm thiểu dữ liệu tham số**

- Có thể được sử dụng để tính toán xấp xỉ các dữ liệu đã cho.

- Trong hồi quy tuyến tính, dữ liệu được mô hình hóa theo một đường thẳng. Biến y (biến phản hồi), có thể được mô hình hóa như một hàm tuyến tính của một biến khác biến ngẫu nhiên, x (được gọi là biến dự đoán), với phương trình

y = ax + b

( Phương sai của y được giả định là hằng số. x và y là các thuộc tính cơ sở dữ liệu số. Các hệ số w và b được gọi là hệ số hồi quy)

- Log-linear là một phần mở rộng của hồi quy tuyến tính (cơ bản), cho phép một biến phản hồi, y được mô hình hóa như một hàm tuyến tính của hai hoặc nhiều biến dự đoán

- Các mô hình log-linear xấp xỉ các phân bố xác suất đa chiều rời rạc.

- Mô hình log-linear có thể được sử dụng để ước tính xác suất của mỗi điểm trong một không gian đa chiều cho một tập hợp các thuộc tính rời rạc, dựa trên một tập hợp con nhỏ hơn của các kết hợp chiều.

- Điều này cho phép một không gian dữ liệu chiều cao hơn được xây dựng từ các không gian chiều thấp hơn. Do đó, các mô hình log-linear cũng hữu ích cho việc giảm tính chiều hướng và làm mịn dữ liệu

Kết luận:

Cả hai mô hình hồi quy và log-linear đều có thể được sử dụng trên dữ liệu thưa thớt, mặc dù chúng ứng dụng có thể bị hạn chế. Trong khi cả hai phương pháp đều có thể xử lý dữ liệu bị lệch, hồi quy làm đặc biệt tốt.

Một số phần mềm tồn tại để giải quyết các vấn đề hồi quy. Ví dụ: e SAS (www.sas.com), SPSS (www.spss.com), and S-Plus (www.insightful.com). Một nguồn hữu ích khác là cuốn sách Numerical Recipes in C, của Press, Teukolsky, Vetterling, and Flannery [PTVF07].

**3.4.6. Biểu đồ tần suất**

Biểu đồ tần suất sử dụng binning để phân phối dữ liệu gần đúng và là một dạng phổ biến giảm dữ liệu. Biểu đồ đã được giới thiệu trong Phần 2.2.3. Biểu đồ tần suất cho một thuộc tính, A, phân vùng phân phối dữ liệu của A thành các tập con rời rạc, được gọi là buckets hoặc bins.

- Nếu mỗi vùng lưu trữ chỉ đại diện cho một cặp giá trị/tần suất thuộc tính duy nhất, thì biểu tượng buckets được gọi là singleton buckets.

Ví dụ 3.3: Biểu đồ tần suất. Dữ liệu sau đây là danh sách giá AllElectronics thường được bán các mặt hàng (làm tròn đến đô la gần nhất). Các số đã được sắp xếp: 1, 1, 5, 5, 5,5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,

18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,

30, 30.

3.7 Biểu đồ tần suất cho giá sử dụng các singleton buckets — mỗi vùng lưu trữ đại diện cho một giá–giá trị/ cặp tần số.

3.8 Biểu đồ tần suất có chiều rộng bằng nhau cho giá, trong đó các giá trị được tổng hợp để mỗi vùng lưu trữ có một chiều rộng đồng nhất là 10$.

Câu hỏi được đặt ra :

"Các vùng lưu trữ được xác định như thế nào và các giá trị thuộc tính được phân vùng như thế nào?"

Trả lời

Có một số quy tắc phân vùng, bao gồm những điều sau đây:

- Chiều rộng bằng nhau: Trong biểu đồ chiều rộng bằng nhau, chiều rộng của mỗi phạm vi vùng lưu trữ là đồng phục

- Tần số bằng nhau (hoặc độ sâu bằng nhau): Trong biểu đồ tần số bằng nhau, các vùng lưu trữ là được tạo ra sao cho, đại khái, tần suất của mỗi vùng lưu trữ là không đổi

Tác dụng của biểu đồ :

- Biểu đồ có hiệu quả cao trong việc xấp xỉ cả dữ liệu thưa thớt và dày đặc, như cũng như dữ liệu có độ lệch cao và thống nhất. Các biểu đồ được mô tả trước đây cho đĩa đơn thuộc tính có thể được mở rộng cho nhiều thuộc tính. Biểu đồ đa chiều có thể nắm bắt sự phụ thuộc giữa các thuộc tính. Những biểu đồ này đã được tìm thấy có hiệu quả trong xấp xỉ dữ liệu với tối đa năm thuộc tính. Cần có thêm các nghiên cứu liên quan đến hiệu quả của biểu đồ đa chiều cho tính chiều cao.

- Singleton buckets rất hữu ích để lưu trữ các ngoại lệ tần số cao.

**3.4.7. Phân cụm**

- Kỹ thuật phân cụm coi các bộ dữ liệu là các đối tượng. Chúng phân vùng các đối tượng thành các nhóm hoặc cụm để các đối tượng trong một cụm là “tương tự” với nhau và “không giống” với các đối tượng trong các cụm khác.

- Sự giống nhau thường được định nghĩa về mức độ “gần” của các đối tượng trong không gian, dựa trên một hàm khoảng cách.

- “Chất lượng” của một cụm có thể được biểu thị bằng đường kính của nó, khoảng cách tối đa giữa hai đối tượng bất kỳ trong cụm.

- Khoảng cách trung tâm là một thước đo thay thế cho chất lượng cụm và được định nghĩa là khoảng cách trung bình của mỗi đối tượng trong cụm từ trung tâm cụm (biểu thị “đối tượng trung bình” hoặc điểm trung bình trong không gian cho cụm)

- Hình 3.3 cho thấy một biểu đồ 2-D của dữ liệu khách hàng liên quan đến các vị trí của khách hàng trong một thành phố. Ba cụm dữ liệu có thể nhìn thấy. Trong giảm dữ liệu, các đại diện cụm của dữ liệu được sử dụng để thay thế dữ liệu thực tế. Hiệu quả của kỹ thuật này phụ thuộc vào dữ liệu của thiên nhiên. Nó hiệu quả hơn nhiều đối với dữ liệu có thể được tổ chức thành các cụm riêng biệt hơn là đối với dữ liệu bị bôi nhọ

**3.4.8 Sampling (Lấy mẫu)**

- Lấy mẫu (Sampling) có thể được sử dụng như một kỹ thuật giảm dữ liệu vì nó cho phép một tập dữ liệu lớn được biểu diễn bằng một mẫu dữ liệu ngẫu nhiên nhỏ hơn nhiều (hoặc tập hợp con).

- Lấy mẫu ngẫu nhiên không thay thế (SRSWOR) (Simple random sample without replacement (SRSWOR) ): là một phương pháp chọn n đơn vị trong số N lần lượt sao cho ở bất kỳ giai đoạn nào của lựa chọn, bất kỳ đơn vị nào trong số các đơn vị còn lại đều có cơ hội được chọn như nhau, tức là 1 / N

VD: Khi một ví dụ (bản ghi) được lấy mẫu, nó sẽ được loại khỏi tập dữ

liệu ban đầu (sẽ không thể được chọn thêm một lần nào nữa)

- Lấy mẫu ngẫu nhiên có thay thế(Simple random sampling with replacement (SRSWR) ) : là phương pháp chọn lần lượt n đơn vị trong số N đơn vị sao cho ở mỗi giai đoạn lựa chọn, mỗi đơn vị đều có cơ hội được chọn như nhau, tức là 1 / N

VD: Khi một ví dụ (bản ghi) được lấy mẫu, nó không bị loại khỏi tập

dữ liệu ban đầu (có thể được chọn nhiều hơn một lần)

- Lấy mẫu cả khối (cluster sampling): là phương pháp tổ chức chọn mẫu, trong đó số đơn vị mẫu được chọn không phải là đơn lẻ mà cùng một lúc chọn ra một khối đơn vị.

- Tiến hành chọn mẫu cả khối:

o Trong phương pháp này, trước hết tổng thể chung được chia thành các khối (như làng, xã, phường, lượng sản phẩm sản xuất trong 1 khoảng thời gian…). Sau đó, ta chọn ngẫu nhiên một số khối và điều tra tất cả các đơn vị trong khối đã chọn.

o Thường dùng phương pháp này khi không có sẵn danh sách đầy đủ của các đơn vị trong tổng thể cần nghiên cứu. Chẳng hạn điều tra tình hình học tín chỉ của sinh viên, tổng thể sinh viên được sắp xếp thành các lớp tín chỉ chứ không lập danh sách sinh viên và có thể chọn ngẫu nhiên một số lớp để điều tra toàn bộ.

- Ví dụ: Chọn mẫu điều tra cá nhân hộ gia đình

o Bước 1: lựa chọn ngẫu nhiên một số huyện.

o Bước 2: các hộ gia đình được lựa chọn trong các huyện vừa được chọn.

o Bước 3: Tất cả cá nhân từ hộ được chọn.

- Lấy mẫu phân tầng (Statified sampling): là phương pháp mà các đơn vị mẫu được chọn khi tổng thể chung đã được phân chia thành các tầng theo tiêu chuẩn liên quan trực tiếp đến mục đích nghiên cứu. Việc chọn các đơn vị từ các tổ được tiến hành theo phương pháp chọn ngẫu nhiên.

->

- Một ưu điểm của việc lấy mẫu để giảm dữ liệu là chi phí lấy mẫu tỷ lệ với kích thước của mẫu. Do đó, độ phức tạp của việc lấy mẫu có khả năng ảnh hưởng đến kích thước của dữ liệu. Đối với kích thước mẫu cố định, độ phức tạp của việc lấy mẫu chỉ tăng tuyến tính khi số thứ nguyên dữ liệu, n, tăng lên, trong khi các kỹ thuật sử dụng biểu đồ, chẳng hạn, tăng theo cấp số nhân ở n.

- Khi được áp dụng để giảm dữ liệu, lấy mẫu thường được sử dụng nhiều nhất để ước tính kết quả cho một truy vấn tổng hợp. Có thể (sử dụng định lý giới hạn trung tâm) để xác định cỡ mẫu đủ để ước lượng một hàm đã cho trong một mức độ sai số xác định. Kích thước mẫu này có thể cực kỳ nhỏ so với kích thước của tập dữ liệu. Lấy mẫu là một lựa chọn tự nhiên để cải tiến dần tập hợp dữ liệu đã giảm. Một tập hợp như vậy có thể được tinh chỉnh thêm bằng cách đơn giản là tăng kích thước mẫu.

**3.4.9. Data Cube Aggregation**

- Hãy tưởng tượng rằng bạn đã thu thập dữ liệu để phân tích. Dữ liệu này bao gồm doanh số bán hàng của AllElectronics mỗi quý, trong các năm 2008 đến 2010. Tuy nhiên, bạn quan tâm đến doanh số hàng năm (tổng mỗi năm), hơn là tổng mỗi quý. Do đó, dữ liệu có thể được tổng hợp để dữ liệu kết quả tóm tắt tổng doanh số bán hàng mỗi năm thay vì mỗi quý. Sự tổng hợp này được minh họa trong Hình 3.10. Tập dữ liệu kết quả có khối lượng nhỏ hơn, không làm mất thông tin cần thiết cho nhiệm vụ phân tích.

Hình 3.10: Dữ liệu bán hàng cho một nhánh nhất định của AllElectronics trong các năm từ 2008 đến 2010. Ở bên trái, doanh số bán hàng được hiển thị theo quý. Ở bên phải, dữ liệu được tổng hợp để cung cấp doanh số hàng năm.

Hình 3.11 Một khối dữ liệu bán hàng tại AllElectronics ..

- Các khối dữ liệu lưu trữ thông tin tổng hợp đa chiều.

- Ví dụ, Hình 3.11 cho thấy một khối dữ liệu để phân tích đa chiều dữ liệu bán hàng liên quan đến doanh số hàng năm trên mỗi loại mặt hàng cho mỗi chi nhánh của AllElectronics.

- Mỗi ô chứa một giá trị dữ liệu tổng hợp, tương ứng với điểm dữ liệu trong không gian đa chiều. (Để dễ đọc, chỉ một số giá trị ô được hiển thị.)

- Khái niệm phân cấp có thể tồn tại cho mỗi thuộc tính, cho phép phân tích dữ liệu ở nhiều mức trừu tượng.

- Ví dụ, một hệ thống phân cấp cho chi nhánh có thể cho phép các chi nhánh được nhóm thành các vùng, dựa trên địa chỉ của chúng.

- Các khối dữ liệu cung cấp khả năng truy cập nhanh vào dữ liệu tóm tắt, được tính toán trước, do đó mang lại lợi ích cho quá trình xử lý phân tích trực tuyến cũng như khai thác dữ liệu.

- Khối được tạo ra ở mức trừu tượng thấp nhất được gọi là hình khối cơ sở(base cuboid).

- Khối cơ sở phải tương ứng với một thực thể như bán hàng hoặc khách hàng. Nói cách khác, mức thấp nhất phải có thể sử dụng được hoặc hữu ích cho việc phân tích.

- Một khối lập phương ở mức trừu tượng cao nhất là khối chóp.

- Đối với dữ liệu bán hàng trong Hình 3.11, khối chóp sẽ cho một tổng số — tổng doanh số bán hàng trong cả ba năm, cho tất cả các loại mặt hàng và cho tất cả các chi nhánh.

- Các khối dữ liệu được tạo cho các mức độ trừu tượng khác nhau thường được gọi là khối lập phương, do đó khối dữ liệu có thể đề cập đến một mạng các khối lập phương thay thế.

- Mỗi mức trừu tượng cao hơn sẽ làm giảm thêm kích thước dữ liệu kết quả. Khi trả lời các yêu cầu khai thác dữ liệu, nên sử dụng khối lập phương nhỏ nhất có liên quan đến nhiệm vụ đã cho.