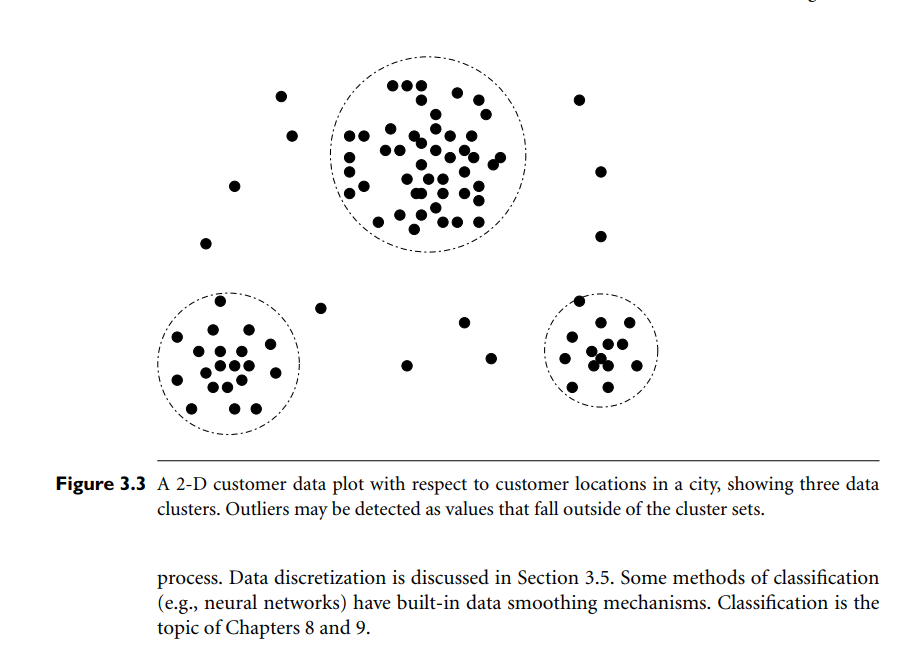
**3.4.7. Phân cụm**

* 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

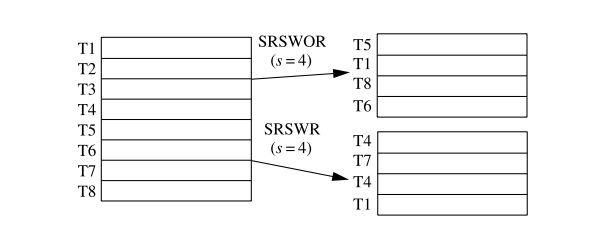
**VD: Khi một ví dụ (bản ghi) được lấy mẫu, nó sẽ được loại khỏi tập dữ**

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

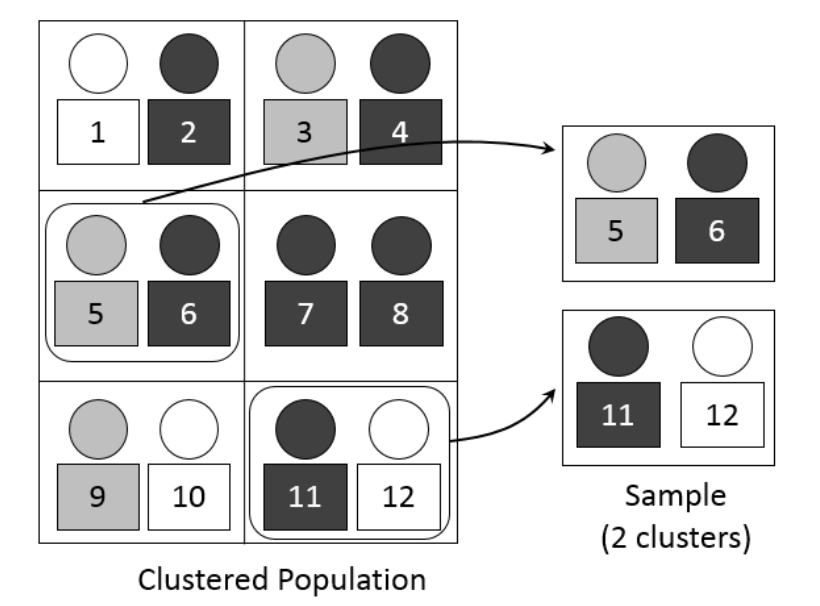
- Lấy mẫu ngẫu nhiên có thay thế(**Simple random sampling with replacement (SRSWR)** ) : là phương pháp chọn lần lượt n đơn vị trong số N đơn vị sao cho ở mỗi giai đoạn lựa chọn, mỗi đơn vị đều có cơ hội được chọn như nhau, tức là 1 / N

**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**:
  + 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.
  + 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
  + Bước 1: lựa chọn ngẫu nhiên một số huyện.
  + Bước 2: các hộ gia đình được lựa chọn trong các huyện vừa được chọn.
  + 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.

Table

Description automatically generated

->

* 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.

Diagram

Description automatically generated

*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.*

Diagram

Description automatically generated

*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.