MỤC LỤC

DANH MỤC CÁC HÌNH MINH HỌA

MỞ ĐẦU

Trong vài thập niên gần đây, cùng với sự thay đổi và phát triển không ngừng của ngành công nghệ thông tin nói chung và trong các ngành công nghệ phần cứng, phân mềm, truyền thông và hệ thống các dữ liệu phục vụ trong các lãnh vực kinh tế - xã hội nói riêng. Thì việc thu thập thông tin cũng như nhu cầu lưu trữ thông tin càng ngày càng lớn. Bên cạnh đó việc tin học hoá một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Hàng triệu Cơ sở dữ liệu đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lí ... trong đó có nhiều Cơ sở dữ liệu cực lớn cỡ Gigabyte, thậm chí là Terabyte. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kĩ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kĩ thuật Khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền Công nghệ thông tin thế giới hiện nay. Một vấn đề được đặt ra là phải làm sao trích chọn được những thông tin có ý nghĩa từ tập dữ liệu lớn để từ đó có thể giải quyết được các yêu cầu của thực tế như trợ giúp ra quyết định, dự đoán,… và Khai phá dữ liệu (Data mining) đã ra đời nhằm giải quyết các yêu cầu đó.

Khai phá dữ liệu được định nghĩa là: quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu…Hiện nay, ngoài thuật ngữ khai phá dữ liệu, người ta còn dùng một số thuật ngữ khác có ý nghĩa tương tự như: khai phá tri thức từ cơ sở dữ liệu (knowlegde mining from databases), trích lọc dữ liệu (knowlegde extraction), phân tích dữ liệu/mẫu (data/pattern analysis), khảo cổ dữ liệu (data archaeology), nạo vét dữ liệu (data dredging). Nhiều người coi khai phá dữ liệu và một thuật ngữ thông dụng khác là khám phá tri thức trong cơ sở dữ liệu (Knowlegde Discovery in Databases – KDD) là như nhau. Tuy nhiên trên thực tế, khai phá dữ liệu chỉ là một bước thiết yếu trong quá trình Khám phá tri thức trong cơ sở dữ liệu.

Ngay từ những ngày đầu khi xuất hiện, Data mining đã trở thành một trong những xu hướng nghiên cứu phổ biến trong lĩnh vực học máy tính và công nghệ tri thức. Nhiều thành tựu nghiên cứu của Data mining đã được áp dụng trong thực tế. Data mining có nhiều hướng quan trọng và một trong các hướng đó là phân cụm dữ liệu (Data Clustering). Phân cụm dữ liệu là quá trính tìm kiếm để phân ra các cụm dữ liệu, các mẫu dữ liệu từ tập Cơ sở dữ liệu lớn. Phân cụm dữ liệu là một phương pháp học không giám sát

Phân cụm dữ liệu là một trong những kỹ thuật để khai thác dữ liệu có hiệu quả. Phân cụm dữ liệu đã được ứng dụng trong nhiều lĩnh vực khác nhau: kinh tế, bảo hiểm, quy hoạch đô thị, nghiên cứu về địa chấn v.v… Tuy nhiên, trong lãnh vực giáo dục, mặc dù là ngành có khối lượng dữ liệu khá lớn, cần phân tích để đưa ra các chiến lược phát triển phù hợp thì thực sự chưa được khai thác có hiệu quả. Bản thân người thực hiện đề tài đang công tác trong ngành giáo dục (ở cấp độ sở), nên rất cần các phân tích, đánh giá kết quả học tập của học sinh để từ đó đề xuất các biện pháp nhằm nâng cao chất lượng giáo dục học sinh phổ thông. Đó là lý do chọn đề tài *"Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh"*.

Bố cục luận văn

Ngoài các phần Mở đầu, Mục lục, Danh mục hình, Kết luận, Tài liệu tham khảo. Luận văn chia là 3 phần :

- Phần 1 : Tổng quan về khám phá tri thức và khai phá dữ liệu

Phần này giới thiệu một cách tổng quát về quá trình khám phá tri thức nói chung và khai phá dữ liệu nói riêng. Các phương pháp, lãnh vực và các hướng tiếp cận trong khai phá dữ liệu.

- Phần 2 : Phân cụm dữ liệu và một số thuật toán trong phân cụm dữ liệu

Trong phần này trình bày khái niệm và mục tiêu của phân cụm dữ liệu, các yêu cầu, các cách tiếp cận cũng như các thách thức mà phân cụm dữ liệu đang gặp phải.

Một số phương pháp phân cụm dữ liệu như: phân cụm không phân cấp, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dữ liệu dựa vào lưới, phân cụm dựa trên mô hình … trong mỗi phương pháp trình bày một số thuật toán đại diện.

- Phần 3 : Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh

Phần này trình bày lý do chọn bài toán, các cơ sở để giải quyết bài toán (lý luận, thực tiễn, khoa học …). Cài đặt chương trình thử nghiệm ứng dụng kỹ thuật phân cụm trong lãnh vực giáo dục và một số kết quả thu được.

Chương 1. Tổng quan về khám phá tri thức và khai phá dữ liệu

1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin và ngành công nghiệp phần cứng đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Bên cạnh đó việc tin học hoá một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lãnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Hàng triệu cơ sở dữ liệu đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lí..., trong đó có nhiều cơ sở dữ liệu cực lớn cỡ Gigabyte, thậm chí là Terabyte. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lãnh vực thời sự của ngành công nghệ thông tin thế giới hiện nay.

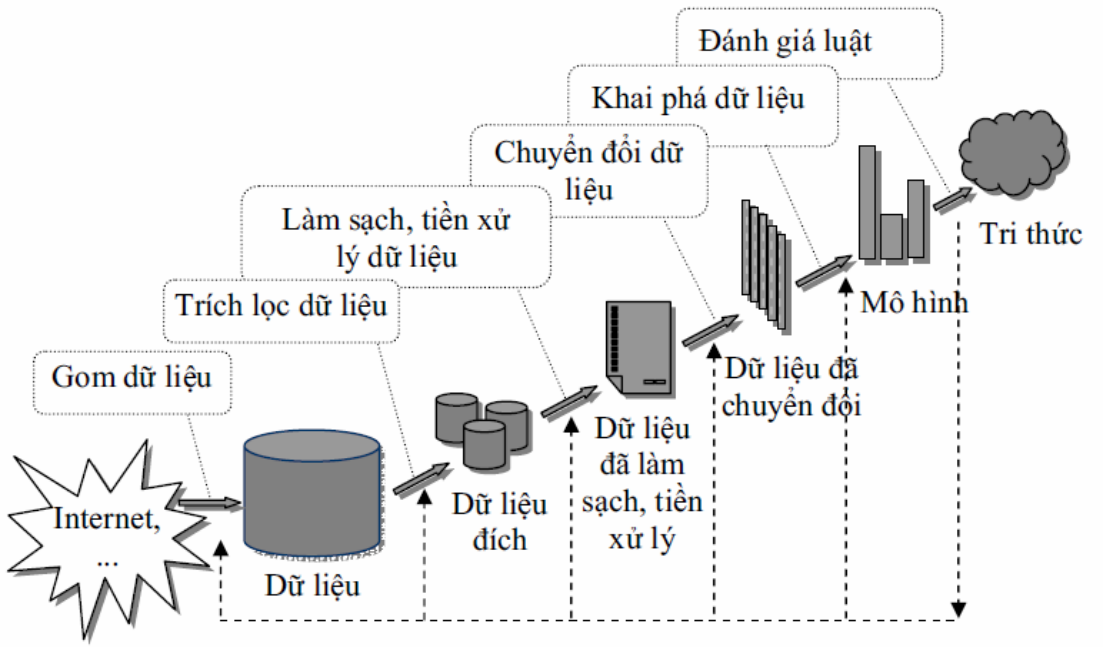
Thông thường, chúng ta coi dữ liệu như là một chuỗi các bits, hoặc các số và các ký hiệu hay là các "đối tượng" với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Các bits thường được sử dụng để đo thông tin, và xem nó như là dữ liệu đã được loại bỏ phần tử thừa, lặp lại, và rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Tri thức được xem như là các thông tin tích hợp, bao gồm các sự kiện và mối quan hệ giữa chúng, đã được nhận thức, khám phá, hoặc nghiên cứu. Nói cách khác, tri thức có thể được coi là dữ liệu ở mức độ cao của sự trừu tượng và tổng quát.

Khám phá tri thức hay phát hiện tri thức trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: Phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của khai phá dữ liệu là tìm kiếm các mẫu hoặc mô hình tồn tại trong cơ sở dữ liệu nhưng ẩn trong khối lượng lớn dữ liệu.

1.2. Quá trình khám phá tri thức

Quá trình khám phá tri thức tiến hành qua 6 giai đoạn như hình [7]:



Hình 1.1 : Quá trình khám phá tri thức

Bắt đầu của quá trình là kho dữ liệu thô và kết thúc với tri thức được chiết xuất ra. Về lý thuyết thì có vẻ rất đơn giản nhưng thực sự đây là một quá trình rất khó khăn gặp phải rất nhiều vướng mắc như : quản lý các tập dữ liệu, phải lặp đi lặp lại toàn bộ quá trình, v.v...

1. Gom dữ liệu: Tập hợp dữ liệu là bước đầu tiên trong quá trình khai phá dữ liệu. Đây là bước được khai thác trong một cơ sở dữ liệu, một kho dữ liệu và thậm chí các dữ liệu từ các nguồn ứng

dụng Web.

2. Trích lọc dữ liệu: Ở giai đoạn này dữ liệu được lựa chọn hoặc phân chia theo một số tiêu chuẩn nào đó phục vụ mục đích khai thác, ví dụ chọn tất cả những em học sinh có điểm Trung bình học kỳ lớn hơn 8.0 và có giới tính nữ.

3. Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu: Giai đoạn thứ ba này là giai đoạn hay bị sao lãng, nhưng thực tế nó là một bước rất quan trọng trong quá trình khai phá dữ liệu. Một số lỗi thường mắc phải trong khi gom dữ liệu là tính không đủ chặt chẽ, logíc. Vì vậy, dữ liệu thường chứa các giá trị vô nghĩa và không có khả năng kết nối dữ liệu. Ví dụ : Điểm Trung bình = 12.4. Giai đoạn này sẽ tiến hành xử lý những dạng dữ liệu không chặt chẻ nói trên. Những dữ liệu dạng này được xem như thông tin dư thừa, không có giá trị. Bởi vậy, đây là một quá trình rất quan trọng vì dữ liệu này nếu không được "làm sạch – tiền xử lý – chuẩn bị trước" thì sẽ gây nên những kết quả sai lệch nghiêm trọng.

4. Chuyển đổi dữ liệu: Tiếp theo là giai đoạn chuyển đổi dữ liệu, dữ liệu đưa ra có thể sử dụng và điều khiển được bởi việc tổ chức lại nó, tức là dữ liệu sẽ được chuyển đổi về dạng phù hợp cho việc khai phá bằng cách thực hiện các thao tác nhóm hoặc tập hợp.

5. Khai phá dữ liệu: Đây là bước mang tính tư duy trong khai phá dữ liệu. Ở giai đoạn này nhiều thuật toán khác nhau đã được sử dụng để trích ra các mẫu từ dữ liệu. Thuật toán thường dùng là nguyên tắc phân loại, nguyên tắc kết, v.v...

6. Đánh giá các luật và biểu diễn tri thức: Ở giai đoạn này, các mẫu dữ liệu được chiết xuất ra bởi phần mềm khai phá dữ liệu. Không phải bất cứ mẫu dữ liệu nào cũng đều hữu ích, đôi khi nó còn bị sai lệch. Vì vậy, cần phải ưu tiên những tiêu chuẩn đánh giá để chiết xuất ra các tri thức (Knowlege) cần chiết xuất ra. Đánh giá sự hữu ích của các mẫu biểu diễn tri thức dựa trên một số phép

đo. Sau đó sử dụng các kỹ thuật trình diễn và trực quan hoá dữ liệu để biểu diễn tri thức khai phá được cho người sử dụng.

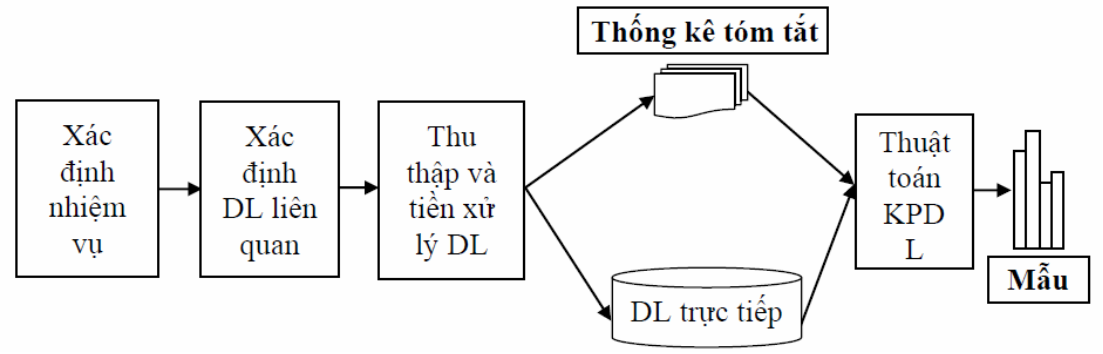
Trên đây là 6 giai đoạn của quá trình khám phá tri thức, trong đó giai đoạn 5 - khai phá dữ liệu (hay còn gọi đó là Data Mining) là giai đoạn được quan tâm nhiều nhất.

1.3. Quá trình khai phá dữ liệu

Khai phá dữ liệu là một giai đoạn quan trọng trong quá trình khám phá tri thức. Về bản chất là giai đoạn duy nhất tìm ra được thông tin mới, thông tin tiềm ẩn có trong cơ sở dữ liệu chủ yếu phục vụ cho mô tả và dự đoán.

Mô tả dữ liệu là tổng kết hoặc diễn tả những đặc điểm chung của những thuộc tính dữ liệu trong kho dữ liệu mà con người có thể hiểu được.

Dự đoán là dựa trên những dữ liệu hiện thời để dự đoán những quy luật được phát hiện từ các mối liên hệ giữa các thuộc tính của dữ liệu trên cơ sở đó chiết xuất ra các mẫu, dự đoán được những giá trị chưa biết hoặc những giá trị tương lai của các biến quan tâm.



Hình 1.2 : Quá trình khai phá dữ liệu

- Xác định nhiệm vụ: Xác định chính xác các vấn đề cần giải quyết.
- Xác định các dữ liệu liên quan: Dùng để xây dựng giải pháp.

- Thu thập và tiền xử lý dữ liệu: Thu thập các dữ liệu liên quan và tiền xử lý chúng sao cho thuật toán khai phá dữ liệu có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều các vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), v.v..
- Thuật toán khai phá dữ liệu: Lựa chọn thuật toán khai phá dữ liệu và thực hiện việc khai phá dữ liệu để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định... tương ứng với ý nghĩa của nó.

1.4. Các phương pháp khai phá dữ liệu

Với hai mục đích khai phá dữ liệu là Mô tả và Dự đoán, người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu [3]:

- Luật kết hợp (association rules)
- Phân lớp (Classfication)
- Hồi qui (Regression)
- Trực quan hóa (Visualiztion)
- Phân cụm (Clustering)
- Tổng hợp (Summarization)
- Mô hình ràng buộc (Dependency modeling)
- Biểu diễn mô hình (Model Evaluation)
- Phân tích sự phát triển và độ lệch (Evolution and deviation analyst)
- Phương pháp tìm kiếm (Search Method)

Có nhiều phương pháp khai phá dữ liệu được nghiên cứu ở trên, trong đó có ba phương pháp được các nhà nghiên cứu sử dụng nhiều nhất đó là : Luật kết hợp, Phân lớp dữ liệu và Phân cụm dữ liệu.

1.5. Các lĩnh vực ứng dụng thực tiễn của khai phá dữ liệu

Khai phá dữ liệu là một lĩnh vực mới phát triển những thu hút được khá nhiều nhà nghiên cứu nhờ vào những ứng dụng thực tiễn của nó. Sau đây là một số lĩnh vực ứng dụng thực tế điển hình của khai phá dữ liệu :

- Phân tích dữ liệu và hỗ trợ ra quyết định
- Phân lớp văn bản, tóm tắt văn bản, phân lớp các trang Web và phân cụm ảnh màu
- Chuẩn đoán triệu chứng, phương pháp trong điều trị y học
- Tìm kiếm, đối sánh các hệ Gene và thông tin di truyền trong sinh học
- Phân tích tình hình tài chính, thị trường, dự báo gía cổ phiếu trong tài chính, thị trường và chứng khoán
- Phân tích dữ liệu marketing, khách hàng.
- Điều khiển và lập lịch trình
- Bảo hiểm
- Giáo dục.....

1.6. Các hướng tiếp cận cơ bản và kỹ thuật áp dụng trong khai phá dữ liệu

Vấn đề khai phá dữ liệu có thể được phân chia theo lớp các hướng tiếp cận chính sau:

- Phân lớp và dự đoán (classification & prediction): Là quá trình xếp

một đối tượng vào một trong những lớp đã biết trước (ví dụ: phân lớp các bệnh nhân theo dữ liệu hồ sơ bệnh án, phân lớp vùng địa lý theo dữ liệu thời tiết...). Đối với hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơron nhân tạo (neural network),...hay lớp bài toán này còn đươc gọi là học có giám sát - Học có thầy (supervised learning).

- Phân cụm (clustering/segmentation): Sắp xếp các đối tượng theo từng cụm dữ liệu tự nhiên, tức là số lượng và tên cụm chưa được biết trước. Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các cụm khác nhau là nhỏ nhất. Lớp bài toán này còn được gọi là học không giám sát - Học không thầy (unsupervised learning).

- Luật kết hợp (association rules): Là dạng luật biểu diễn tri thức ở dạng khá đơn giản (Ví dụ: 80% sinh viên đăng ký học Cơ sở dữ liệu thì có tới 60% trong số họ đăng ký học Phân tích thiết kế hệ thống thông tin). Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin sinh học, giáo dục, viễn thông, tài chính và thị trường chứng khoán,...

- Phân tích chuỗi theo thời gian (sequential/temporal patterns): Cũng tưng tự như khai phá dữ liệu bằng luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Một luật mô tả mẫu tuần tự có dạng tiêu biểu X -> Y, phản ánh sự xuất hiện của biến cố X sẽ dẫn đến việc xuất hiện biến cố Y. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán bởi chúng có tính dự báo cao.

- Mô tả khái niệm (concept desccription & summarization): Lớp bài toán này thiên về mô tả, tổng hợp và tóm tắt khái niệm (Ví dụ: tóm tắt văn bản).

1.7. Những thách thức - khó khăn trong khám phá tri thức và khai phá dữ liệu

Khám phá tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực trong thực tế, vì vậy các thách thức và khó khăn ngày càng nhiều, càng lớn hơn. Sau đây là một số các thách thức và khó khăn cần được quan tâm [3]:

- Các cơ sở dữ liệu lớn hơn rất nhiều : cơ sở dữ liệu với hàng trăm trường và bảng, hàng triệu bản ghi và kích thước lên tới nhiều gigabyte là vấn đề hoàn toàn bình thường.
- Số chiều cao : không chỉ thường có một số lượng rất lớn các bản ghi trong cơ sở dữ liệu mà còn có một số lượng rất lớn các trường (các thuộc tính, các biến) làm cho số chiều của bài toán trở nên cao. Thêm vào đó, nó tăng thêm cơ hội cho một giải thuật khai phá dữ liệu tìm ra các mẫu không hợp lệ.
- Thay đổi dữ liệu và tri thức : thay đổi nhanh chóng dữ liệu (động) có thể làm cho các mẫu phát hiện trước đó không hợp lệ. Thêm vào đó, các biến đã đo trong một cơ sở dữ liệu ứng dụng cho trước có thể bị sửa đổi, xóa bỏ hay tăng thêm các phép đo mới. Các giải pháp hợp lý bao gồm các phương pháp tăng trưởng để cập nhật các mẫu và xử lý thay đổi.
- Dữ liệu thiếu và bị nhiễu : bài toán này đặc biệt nhạy trong các cơ sở dữ liệu thương mại. Các thuộc tính quan trọng có thể bị mất nếu cơ sở dữ liệu không được thiết kế với sự khám phá bằng trí tuệ. Các giải pháp có thể gồm nhiều chiến lược thống kê phức tạp để nhận biết các biến ẩn và các biến phụ thuộc.
- Mối quan hệ phức tạp giữa các trường : các thuộc tính hay giá các giá trị có cầu trúc phân cấp, các quan hệ giữa các thuộc tính và các

phương tiện tinh vi hơn cho việc biểu diễn tri thức về nội dung của một cơ sở dữ liệu sẽ đòi hỏi các giải thuật phải có khả năng sử dụng hiệu quả các thông tin này. Về mặt lịch sử, các giải thuật khai phá dữ liệu được phát triển cho các bản ghi có giá trị thuộc tính đơn giản, mặc dù các kỹ thuật mới bắt nguồn từ mối quan hệ giữa các biến đang được phát triển.

- Tính dễ hiểu của các mẫu : trong nhiều ứng dụng, điều quan trọng là những gì khai thác được phải càng dễ hiểu đối với con người thì càng tốt. Các giải pháp có thể thực hiện được bao gồm cả việc biểu diễn được minh họa bằng đồ thị, cấu trúc luật với các đồ thị có hướng, biểu diễn bằng ngôn ngữ tự nhiên và các kỹ thuật hình dung ra dữ liệu và tri thức.

- Người dùng tương tác và tri thức sẵn có : nhiều phương pháp khám phá tri thức và các công cụ không tương tác thực sự với người dùng và không thể dễ dàng kết hợp chặt chẽ với tri thức có sẵn về một bài toán loại trừ theo các cách đơn giản. Việc sử dụng của miền tri thức là quan trọng trong toàn bộ các bước của xử lý khám phá tri thức.

- Tích hợp với các hệ thống khác: Một hệ thống phát hiện đứng một mình có thể không hữu ích lắm. Các vấn đề tích hợp điển hình gồm có việc tích hợp với một DBMS (tức là qua một giao diện truy vấn), tích hợp với các bảng tính và các công cụ trực quan và điều tiết các dự đoán cảm biến thời gian thực.

1.8. Kết luận

Khai phá dữ liệu là lĩnh vực đã và đang trở thành một trong những hướng nghiên cứu thu hút được sự quan tâm của nhiều chuyên gia về công nghệ thông tin trên thế giới. Trong những năm gần đây, rất nhiều phương pháp và thuật toán mới liên tục được công bố. Điều này chứng tỏ những ưu

thế, lợi ích và khả năng ứng dụng thực tế to lớn của khai phá dữ liệu. Chương này đã trình bày một số kiến thức tổng quan về khám phá tri thức, những khái niệm và kiến thức cơ bản nhất về khai phá dữ liệu.

Chương 2. Phân cụm dữ liệu và một số phương pháp phân cụm dữ liệu

2.1. Khái niệm và mục tiêu của phân cụm dữ liệu

2.1.1. Phân cụm dữ liệu là gì ?

Phân cụm dữ liệu là một kỹ thuật trong Data mining nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn và quan trọng trong tập dữ liệu lớn để từ đó cung cấp thông tin, tri thức cho việc ra quyết định.

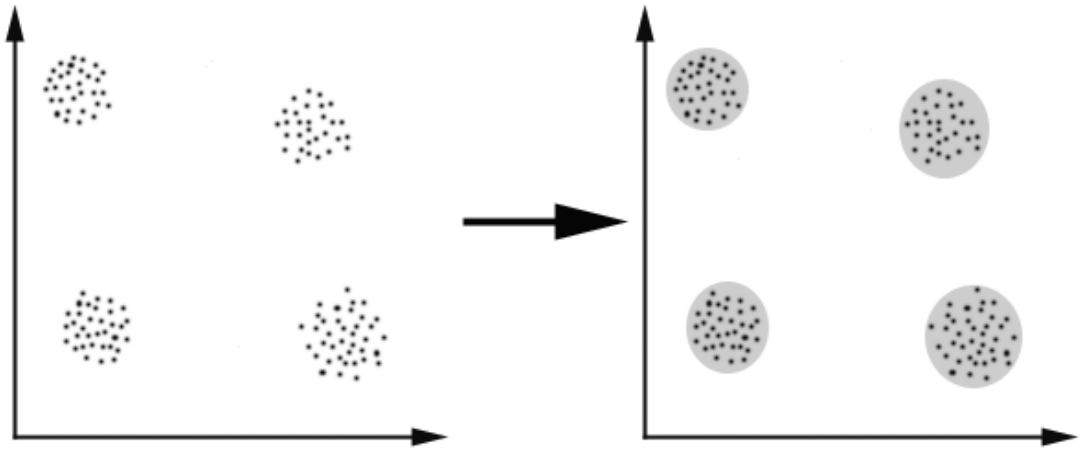
Phân cụm dữ liệu là sự phân chia một cơ sở dữ liệu lớn thành các nhóm dữ liệu với trong đó các đối tượng tương tự như nhau. Trong mỗi nhóm, một số chi tiết có thể không quan tâm đến để đổi lấy dữ liệu đơn giản hóa. Hay ta có thể hiểu "Phân cụm dữ liệu là quá trình tổ chức các đối tượng thành từng nhóm mà các đối tượng ở mỗi nhóm đều tương tự nhau theo một tính chất nào đó, những đối tượng không tương tự tính chất sẽ ở nhóm khác" [1].

Phân cụm dữ liệu là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ liệu là một ví dụ của phương pháp học không có thầy. Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ . . . Ngoài ra phân cụm dữ liệu còn có thể được sử dụng như một bước tiền xử lí cho các thuật toán khai phá dữ liệu khác như là phân loại và mô tả đặc điểm, có tác dụng trong việc phát hiện ra các cụm.

Như vậy, phân cụm dữ liệu là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các đối tượng trong một cụm "tương tự" (Similar) với nhau và các đối tượng trong các cụm khác nhau sẽ "không

tương tự" (Dissimilar) với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định.

Chúng ta có thể thấy điều này với một ví dụ đơn giản như sau [8]:



Hình 2.1: Ví dụ về phân cụm dữ liệu

Trong trường hợp này, chúng ta dễ dàng xác định được 4 cụm dựa vào các dữ liệu đã cho; các tiêu chí "tương tự" để phân cụm trong trường hợp này là khoảng cách : hai hoặc nhiều đối tượng thuộc nhóm của chúng được "đóng gói" theo một khoảng cách nhất định. Điều này được gọi là phân cụm dựa trên khoảng cách.

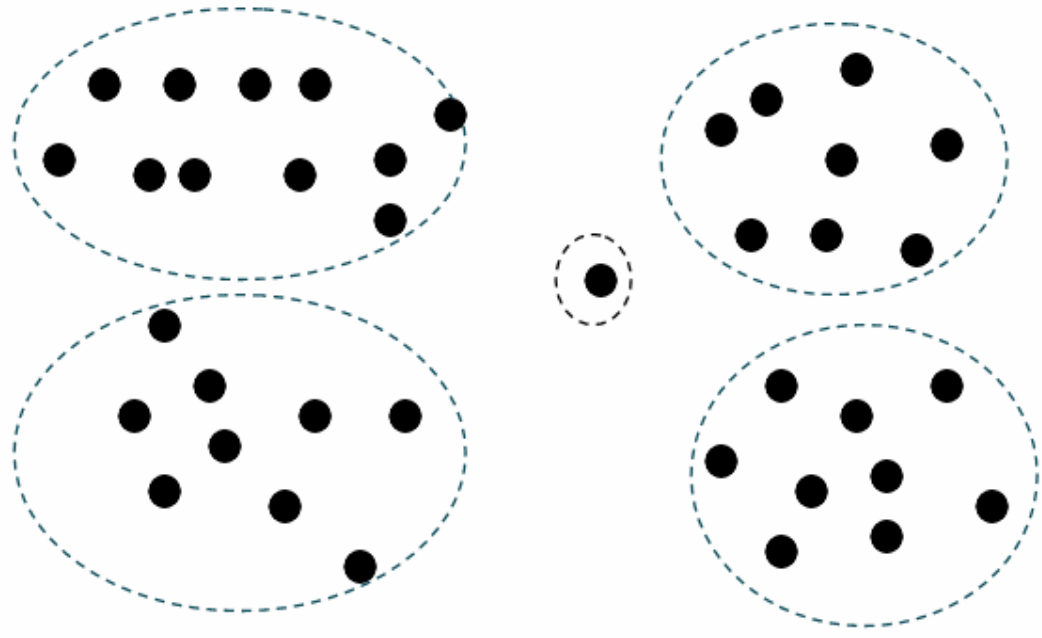
Một kiểu khác của phân cụm dữ liệu là phân cụm dữ liệu dựa vào khái niệm: hai hay nhiều đối tượng thuộc cùng nhóm nếu có một định nghĩa khái niệm chung cho tất cả các đối tượng trong đó. Nói cách khác, đối tượng của nhóm phải phù hợp với nhau theo miêu tả các khái niệm đã được định nghĩa, không phải theo những biện pháp đơn giản tương tự.

2.1.2. Các mục tiêu của phân cụm dữ liệu

Mục tiêu của phân cụm dữ liệu là để xác định các nhóm nội tại bên trong một bộ dữ liệu không có nhãn. Nhưng để có thể quyết định được cái gì

tạo thành một cụm tốt. Nhưng làm thế nào để quyết định cái gì đã tạo nên một phân cụm dữ liệu tốt ? Nó có thể được hiển thị rằng không có tiêu chuẩn tuyệt đối "tốt nhất" mà sẽ là độc lập với mục đích cuối cùng của phân cụm dữ liệu. Do đó, mà người sử dụng phải cung cấp tiêu chuẩn, theo cách như vậy mà kết quả của phân cụm dữ liệu sẽ phù hợp với nhu cầu của họ cần.

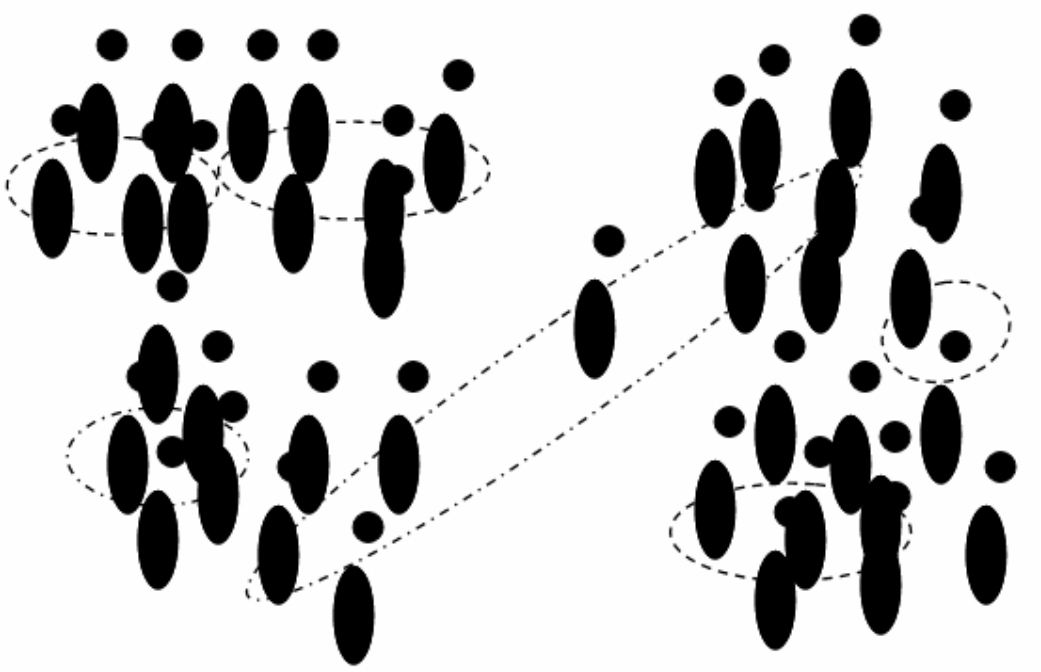
Ví dụ, chúng ta có thể quan tâm đến việc tìm kiếm đối tượng đại diện cho các nhóm đồng nhất trong "các cụm tự nhiên" và mô tả thuộc tính không biết của chúng trong việc tìm kiếm các nhóm hữu ích và phù hợp hoặc trong việc tìm kiếm các đối tượng bất thường trong dữ liệu (cá biệt, ngoại lệ, nhiễu) [1].



Hình 2.2: Ví dụ phân cụm các ngôi nhà dựa trên khoảng cách

Một vấn đề thường gặp trong phân cụm là hầu hết các dữ liệu cần cho phân cụm đều có chứa dữ liệu nhiễu do quá trình thu thập thiếu chính xác hoặc thiếu đầy đủ, vì vậy cần phải xây dựng chiến lược cho bước tiền xử lí dữ liệu nhằm khắc phục hoặc loại bỏ nhiễu trước khi chuyển sang giai đoạn phân tích cụm dữ liệu. Nhiễu ở đây được hiểu là các đối tượng dữ liệu không chính

xác, không tường minh hoặc là các đối tượng dữ liệu khuyết thiếu thông tin về một số thuộc tính... Một trong các kỹ thuật xử lí nhiễu phổ biến là việc thay thế giá trị các thuộc tính của đối tượng nhiễu bằng giá trị thuộc tính tương ứng. Ngoài ra, dò tìm đối tượng ngoại lai cũng là một trong những hướng nghiên cứu quan trọng trong phân cụm, chức năng của nó là xác định một nhóm nhỏ các đối tượng dữ liệu khác thường so với các dữ liệu trong cơ sở dữ liệu, tức là các đối tượng dữ liệu không tuân theo các hành vi hoặc mô hình dữ liệu nhằm tránh sự ảnh hưởng của chúng tới quá trình và kết quả của phân cụm.



Hình 2.3: Ví dụ phân cụm các ngôi nhà dựa trên kích cở

Theo các nghiên cứu đến thời điểm hiện nay thì chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cơ sở dữ liệu. Hơn nữa, đối với các phương pháp phân cụm cần có cách thức biểu diễn cấu trúc của cơ sở dữ liệu, với mỗi cách thức biểu diễn khác nhau sẽ có tương ứng một thuật toán phân cụm phù hợp. Vì vậy phân cụm dữ liệu vẫn đang là một vấn đề khó và mở, vì phải giải quyết nhiều vấn

đề cơ bản một cách trọn vẹn và phù hợp với nhiều dạng dữ liệu khác nhau, đặc biệt là đối với dữ liệu hỗn hợp đang ngày càng tăng trong các hệ quản trị dữ liệu và đây cũng là một trong những thách thức lớn trong lĩnh vực khai phá dữ liệu.

2.2. Các ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu có thể ứng dụng trong nhiều lãnh vực như [5]:

- Thương mại: tìm kiếm nhóm các khách hàng quan trọng dựa vào các thuộc tính đặc trưng tương đồng và những đặc tả của họ trong các bản ghi mua bán của cơ sở dữ liệu;
- Sinh học: phân loại động, thực vật qua các chức năng gen tương đồng của chúng;
- Thư viện : phân loại các cụm sách có nội dung và ý nghĩa tương đồng nhau để cung cấp cho độc giả, cũng như đặt hàng với nhà cung cấp;
- Bảo hiểm : nhận dạng nhóm tham gia bảo hiểm có chi phí yêu cầu bồi thường trung bình cao, xác định gian lận trong bảo hiểm thông qua các mẫu cá biệt;
- Quy hoạch đô thị : nhận dạng các nhóm nhà theo kiểu, vị trí địa lí, giá trị ... nhằm cung cấp thông tin cho quy hoạch đô thị;
- Nghiên cứu địa chấn : phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho việc nhận dạng các vùng nguy hiểm;
- WWW : tài liệu phân loại, phân nhóm dữ liệu weblog để khám phá các nhóm về các hình thức tiếp cận tương tự trợ giúp cho việc khai phá thông tin từ dữ liệu.

2.3. Các yêu cầu và những vấn đề còn tồn tại trong phân cụm dữ liệu

2.3.1. Các yêu cầu của phân cụm dữ liệu

Phân cụm là một thách thức trong lĩnh vực nghiên cứu ở chỗ những ứng dụng tiềm năng của chúng được đưa ra ngay chính trong những yêu cầu đặc biệt của chúng. Sau đây là những yêu cầu cơ bản của phân cụm trong khai phá dữ liệu:

- Có khả năng mở rộng : nhiều thuật toán phân cụm làm việc tốt với những tập dữ liệu nhỏ chứa ít hơn 200 đối tượng, tuy nhiên, một cơ sở dữ liệu lớn có thể chứa tới hàng triệu đối tượng. Việc phân cụm với một tập dữ liệu lớn có thể làm ảnh hưởng tới kết quả. Vậy làm cách nào để chúng ta có thể phát triển các thuật toán phân cụm có khả năng mở rộng cao đối với các cơ sở dữ liệu lớn ?

- Khả năng thích nghi với các kiểu thuộc tính khác nhau: nhiều thuật toán được thiết kế cho việc phân cụm dữ liệu có kiểu khoảng (kiểu số). Tuy nhiên, nhiều ứng dụng có thể đòi hỏi việc phân cụm với nhiều kiểu dữ liệu khác nhau, như kiểu nhị phân, kiểu tường minh (định danh - không thứ tự), và dữ liệu có thứ tự hay dạng hỗn hợp của những kiểu dữ liệu này.

- Khám phá các cụm với hình dạng bất kỳ: nhiều thuật toán phân cụm xác định các cụm dựa trên các phép đo khoảng cách Euclidean và khoảng cách Manhattan. Các thuật toán dựa trên các phép đo như vậy hướng tới việc tìm kiếm các cụm hình cầu với mật độ và kích cỡ tương tự nhau. Tuy nhiên, một cụm có thể có bất cứ một hình dạng nào. Do đó, việc phát triển các thuật toán có thể khám phá ra các cụm có hình dạng bất kỳ là một việc làm quan trọng.

- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào: nhiều thuật toán phân cụm yêu cầu người dùng đưa vào những tham số nhất định trong phân tích phân cụm (như số lượng các cụm mong muốn).

Kết quả của phân cụm thường khá nhạy cảm với các tham số đầu vào. Nhiều tham số rất khó để xác định, nhất là với các tập dữ liệu có lượng các đối tượng lớn. Điều này không những gây trở ngại cho người dùng mà còn làm cho khó có thể điều chỉnh được chất lượng của phân cụm.

- Khả năng thích nghi với dữ liệu nhiễu: hầu hết những cơ sở dữ liệu thực đều chứa đựng dữ liệu ngoại lai, dữ liệu lỗi, dữ liệu chưa biết hoặc dữ liệu sai. Một số thuật toán phân cụm nhạy cảm với dữ liệu như vậy và có thể dẫn đến chất lượng phân cụm thấp.

- Ít nhạy cảm với thứ tự của các dữ liệu vào: một số thuật toán phân cụm nhạy cảm với thứ tự của dữ liệu vào, ví dụ như với cùng một tập dữ liệu, khi được đưa ra với các thứ tự khác nhau thì với cùng một thuật toán có thể sinh ra các cụm rất khác nhau. Do đó, việc quan trọng là phát triển các thuật toán mà ít nhạy cảm với thứ tự vào của dữ liệu.

- Số chiều lớn: một cơ sở dữ liệu hoặc một kho dữ liệu có thể chứa một số chiều hoặc một số các thuộc tính. Nhiều thuật toán phân cụm áp dụng tốt cho dữ liệu với số chiều thấp, bao gồm chỉ từ hai đến 3 chiều. Người ta đánh giá việc phân cụm là có chất lượng tốt nếu nó áp dụng được cho dữ liệu có từ 3 chiều trở lên. Nó là sự thách thức với các đối tượng dữ liệu cụm trong không gian với số chiều lớn, đặc biệt vì khi xét những không gian với số chiều lớn có thể rất thưa và có độ nghiêng lớn.

- Phân cụm ràng buộc: nhiều ứng dụng thực tế có thể cần thực hiện phân cụm dưới các loại ràng buộc khác nhau. Một nhiệm vụ đặt ra là đi tìm những nhóm dữ liệu có trạng thái phân cụm tốt và thỏa mãn các ràng buộc.

- Dễ hiểu và dễ sử dụng: Người sử dụng có thể chờ đợi những kết quả phân cụm dễ hiểu, dễ lý giải và dễ sử dụng. Nghĩa là, sự phân cụm có

thể cần được giải thích ý nghĩa và ứng dụng rõ ràng.

Với những yêu cầu đáng lưu ý này, nghiên cứu của ta về phân tích phân cụm diễn ra như sau:

- Đầu tiên, ta nghiên cứu các kiểu dữ liệu khác nhau và cách chúng có thể gây ảnh hưởng tới các phương pháp phân cụm.
- Thứ hai, ta đưa ra một cách phân loại chung trong các phương pháp phân cụm.
- Sau đó, ta nghiên cứu chi tiết mỗi phương pháp phân cụm, bao gồm các phương pháp phân hoạch, phân cấp, dựa trên mật độ,... Ta cũng khảo sát sự phân cụm trong không gian đa chiều và các biến thể của các phương pháp khác.

2.3.2. Những vấn đề còn tồn tại trong phân cụm dữ liệu

Có một số vấn đề với phân cụm dữ liệu. Một trong số đó là [5]:

- Kỹ thuật clustering hiện nay không trình bày được tất cả các yêu cầu đầy đủ (và đồng thời);
- Giao dịch với số lượng lớn các mẫu và số lượng lớn các mẫu tin của dữ liệu có thể gặp vấn đề phức tạp về thời gian;
- Hiệu quả của phương pháp phụ thuộc vào định nghĩa của "khoảng cách" (đối với phân cụm dữ liệu dựa trên khoảng cách). Nếu không tồn tại một thước đo khoảng cách rõ ràng chúng ta "phải tự xác định", một điều mà không thật sự dễ dàng chút nào, nhất là trong không gian đa chiều;
- Kết quả của thuật toán phân cụm dữ liệu có thể được giải thích theo nhiều cách khác nhau (mà trong nhiều trường hợp chỉ có thể được giải thích theo ý riêng của mỗi người).

2.4. Những kỹ thuật tiếp cận trong phân cụm dữ liệu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các phương pháp tiếp cận chính như sau : phân cụm phân hoạch (Partitioning Methods); phân cụm phân cấp (Hierarchical Methods); phân cụm dựa trên mật độ (Density-Based Methods); phân cụm dựa trên lưới (Grid-Based Methods); phân cụm dựa trên mô hình phân cụm (Model-Based Clustering Methods) và phân cụm có dữ liệu ràng buộc (Binding data Clustering Methods) [5]

2.4.1. Phương pháp phân cụm phân hoạch (Partitioning Methods)

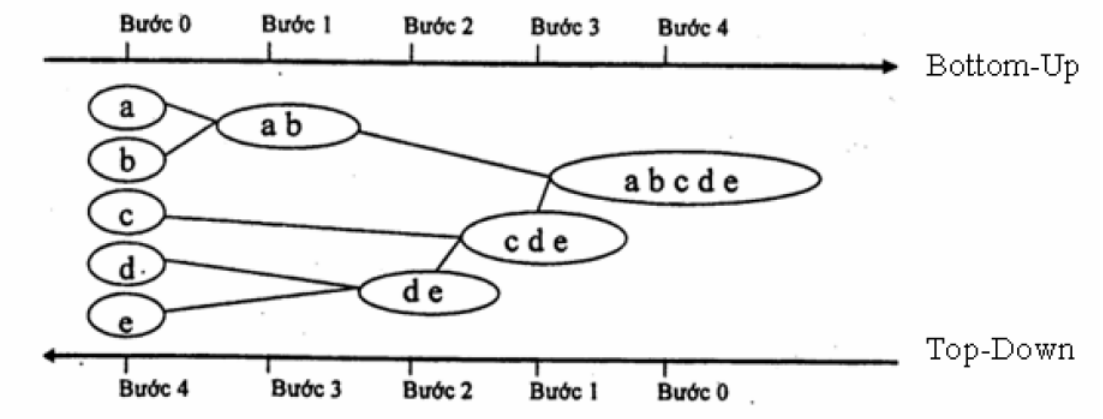
Kỹ thuật này phân hoạch một tập hợp dữ liệu có n phần tử thành k nhóm cho đến khi xác định số các cụm được thiết lập. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước. Phương pháp này là tốt cho việc tìm các cụm hình cầu trong không gian Euclidean. Ngoài ra, phương pháp này cũng phụ thuộc vào khoảng cách cơ bản giữa các điểm để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau so với mỗi điểm khác. Tuy nhiên, phương pháp này không thể xử lí các cụm có hình dạng kỳ quặc hoặc các cụm có mật độ các điểm dầy đặc. Các thuật toán phân hoạch dữ liệu có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, do nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham (Greedy) để tìm kiếm nghiệm.

Điển hình trong phương pháp tiếp cận theo phân cụm phân hoạch là các thuật toán như : K_means, K-medoids, CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based on RAndomized Search) . . .

2.4.2. Phương pháp phân cụm phân cấp (Hierarchical Methods)

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Có hai cách tiếp cận phổ biến của kỹ thuật này đó là: hòa nhập nhóm, thường được gọi là tiếp cận (Bottom-Up); phân chia nhóm, thường được gọi là tiếp cận (Top-Down)

- *Phương pháp "dưới lên" (Bottom up)* : Phương pháp này bắt đầu với mỗi đối tượng được khởi tạo tương ứng với các cụm riêng biệt, sau đó tiến hành nhóm các đối tượng theo một độ đo tương tự (như khoảng cách giữa hai trung tâm của hai nhóm), quá trình này được thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm (mức cao nhất của cây phân cấp) hoặc cho đến khi các điều kiện kết thúc thỏa mãn. Như vậy, cách tiếp cận này sử dụng chiến lược ăn tham trong quá trình phân cụm.



Hình 2.4: Các chiến lược phân cụm phân cấp [7]

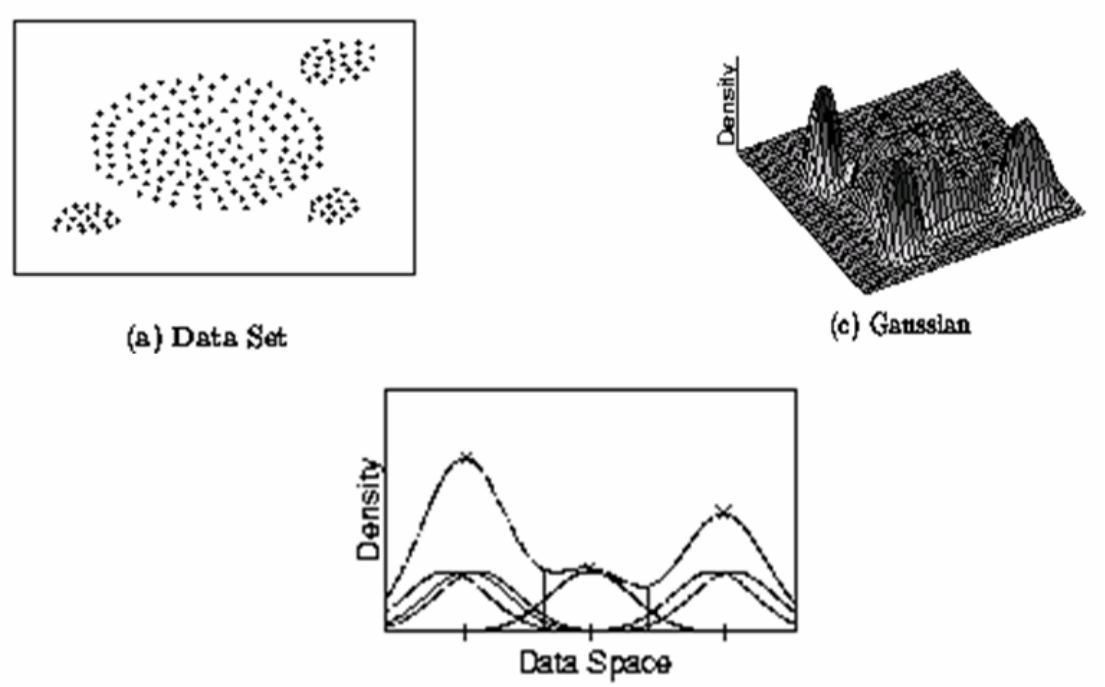
- *Phương pháp "trên xuống" (Top Down)* : Bắt đầu với trạng thái là tất cả các đối tượng được xếp trong cùng một cụm. Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự nào đó cho đến khi mỗi đối tượng là một cụm, hoặc cho đến khi điều kiện dừng thỏa mãn. Cách tiếp cận này sử dụng chiến lược chia để trị trong quá trình phân cụm.

Điển hình trong phương pháp tiếp cận theo phân cụm phân cấp là các thuật toán như : AGNES (Agglomerative Nesting), DIANA (Divisive Analysis), BIRCH (1996), CURE (1998), CHAMELEON (1999) . . .

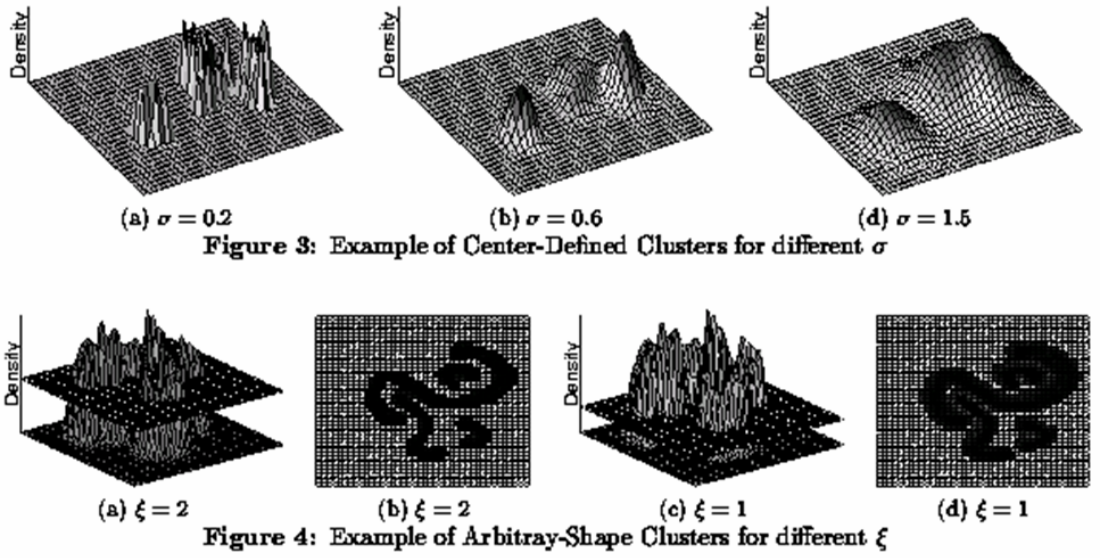
Thực tế áp dụng, có nhiều trường hợp kết hợp cả hai phương pháp phân cụm phân hoạch và phân cụm phân cấp, nghĩa là kết quả thu được của phương pháp phân cấp có thể cải tiến thông qua bước phân cụm phân hoạch. Phân cụm phân hoạch và phân cụm phân cấp là hai phương pháp phân cụm dữ liệu cổ điển, hiện đã có rất nhiều thuật toán cải tiến dựa trên hai phương pháp này đã được áp dụng phổ biến trong khai phá dữ liệu.

2.4.3. Phương pháp phân cụm dựa trên mật độ (Density-Based Methods)

Kỹ thuật này nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định, mật độ là số các đối tượng lân cận của một đối tượng dữ liệu theo một nghĩa nào đó. Trong cách tiếp cận này, khi một dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa trên mật độ của các đối tượng để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Kỹ thuật này có thể khắc phục được các phần tử ngoại lai hoặc giá trị nhiễu rất tốt, tuy nhiên việc xác định các tham số mật độ của thuật toán là rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm.



Hình 2.5: Ví dụ về phân cụm theo mật độ (1) [7]

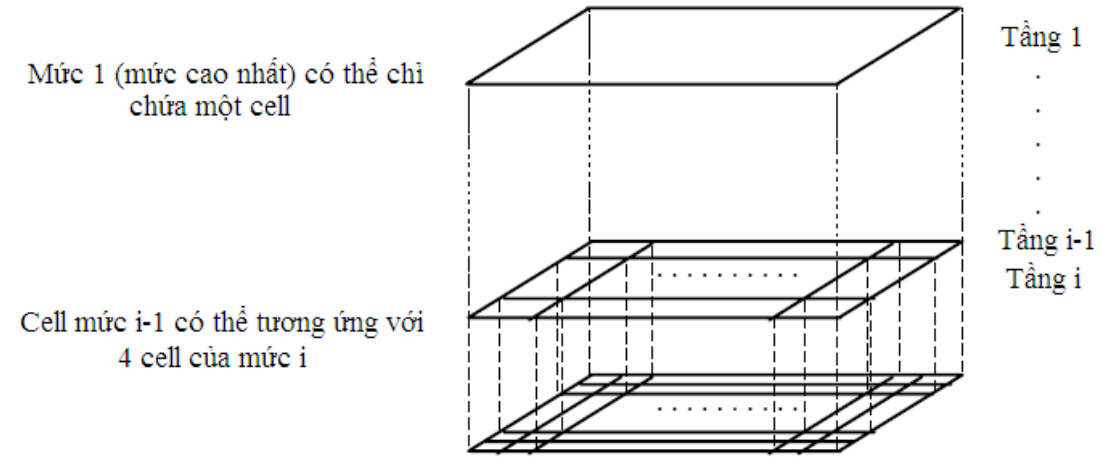


Hình 2.6: Ví dụ về phân cụm theo mật độ (2) [7]

Điển hình trong phương pháp tiếp cận theo phân cụm dựa trên mật độ là các thuật toán như : DBSCAN(KDD'96), DENCLUE (KDD'98), CLIQUE (SIGMOD'98)), OPTICS (SIGMOD'99) . . .

2.4.4. Phương pháp phân cụm dựa trên lưới (Grid-Based Methods)

Kỹ thuật phân cụm dựa trên lưới thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Mục tiêu của phương pháp này là lượng hóa dữ liệu thành các ô tạo thành cấu trúc dữ liệu lưới. Sau đó, các thao tác phân cụm chỉ cần làm việc với các đối tượng trong từng ô trên lưới chứ không phải các đối tượng dữ liệu. Cách tiếp cận dựa trên lưới này không di chuyển các đối tượng trong các ô mà xây dựng nhiều mức phân cấp của nhóm các đối tượng trong một ô. Phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chúng không trộn các ô, đồng thời giải quyết khắc phục yêu cầu đối với dữ liệu nhiều chiều mà phương pháp phân phân cụm dựa trên mật độ không giải quyết được. ưu điểm của phương pháp phân cụm dựa trên lưới là thời gian xử lí nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ô trong mỗi chiều của không gian lưới.



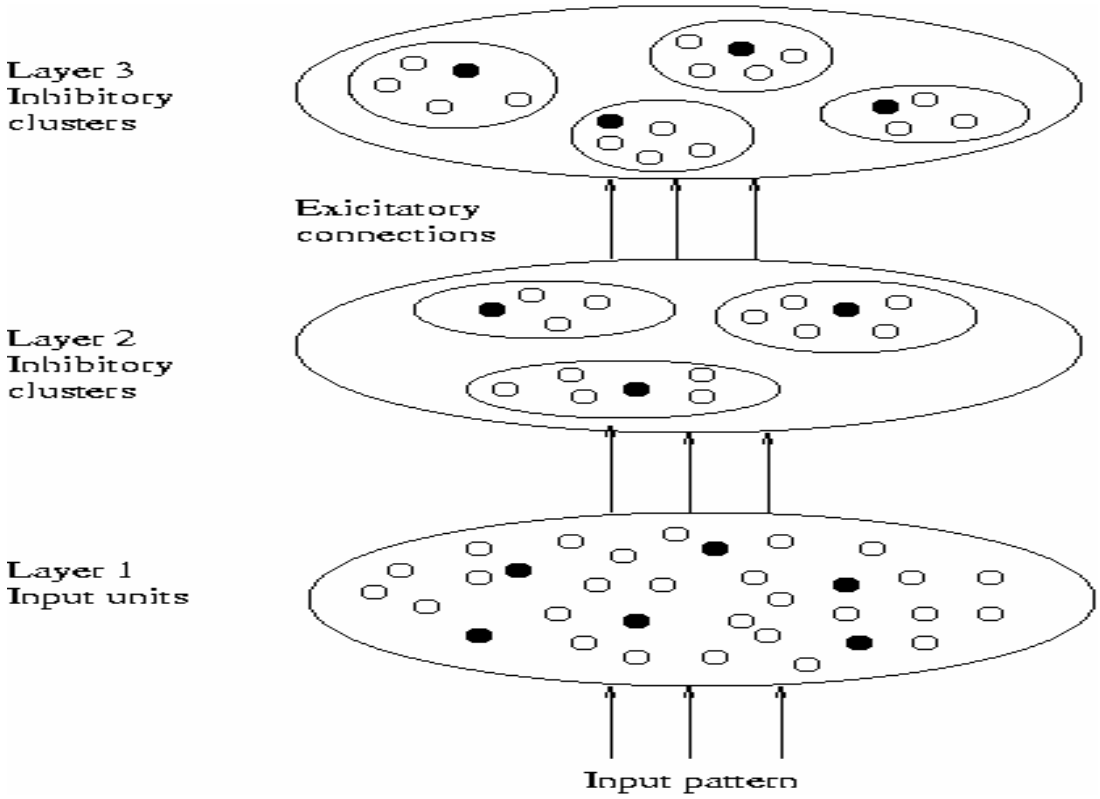
Hình 2.7: Cấu trúc phân cụm dựa trên lưới [7]

Điển hình trong phương pháp tiếp cận theo phân cụm dựa trên lưới là các thuật toán như : STING (a STatistical INformation Grid approach) bởi Wang, Yang và Muntz (1997), WAVECLUSTER bởi Sheikholeslami,

Chatterjee và Zhang (1998), CLIQUE (Clustering In QUEst) bởi Agrawal, Gehrke, Gunopulos, Raghavan (1998) . . .

2.4.5. Phương pháp phân cụm dựa trên mô hình (Model-Based Clustering Methods)

Phương này cố gắng khám phá các phép xấp xỉ tốt của các tham số mô hình sao cho khớp với dữ liệu một cách tốt nhất. Chúng có thể sử dụng chiến lược phân cụm phân hoạch hoặc phân cụm phân cấp, dựa trên cấu trúc hoặc mô hình mà chúng giả định về tập dữ liệu và cách chúng hiệu chỉnh các mô hình này để nhận dạng ra các phân hoạch.



Hình 2.8: Ví dụ về phân cụm dựa trên mô hình [7]

Phương pháp phân cụm dựa trên mô hình cố gắng khớp giữa các dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô

hình có hai cách tiếp cận chính: mô hình thống kê và mạng nơron. Phương pháp này gần giống với phương pháp phân cụm dựa trên mật độ, vì chúng phát triển các cụm riêng biệt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng đôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm.

Điển hình trong phương pháp tiếp cận theo phân cụm dựa trên mô hình là các thuật toán như : EM, COBWEB, CLASSIT, AutoClass (Cheeseman and Stutz, 1996) . . .

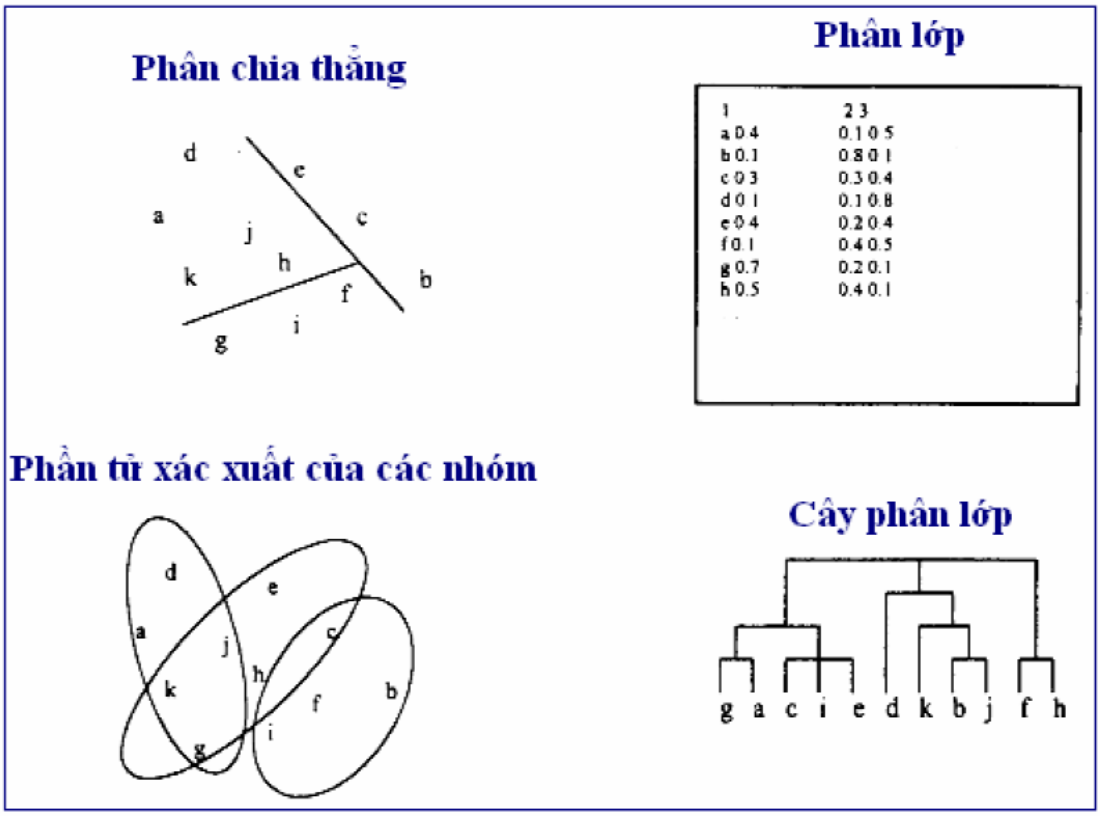
2.4.6. Phương pháp phân cụm có dữ liệu ràng buộc (Binding data Clustering Methods)

Sự phát triển của phân cụm dữ liệu không gian trên cơ sở dữ liệu lớn đã cung cấp nhiều công cụ tiện lợi cho việc phân tích thông tin địa lí, tuy nhiên hầu hết các thuật toán này cung cấp rất ít cách thức cho người dùng để xác định các ràng buộc trong thế giới thực cần phải được thỏa mãn trong quá trình phân cụm. Để phân cụm dữ liệu không gian hiệu quả hơn, các nghiên cứu bổ sung cần được thực hiện để cung cấp cho người dùng khả năng kết hợp các ràng buộc trong thuật toán phân cụm.

Hiện nay, các phương pháp phân cụm trên đã và đang được phát triển và áp dụng nhiều trong các lĩnh vực khác nhau và đã có một số nhánh nghiên cứu được phát triển trên cơ sở của các phương pháp đó như:

- Phân cụm thống kê: Dựa trên các khái niệm phân tích hệ thống, nhánh nghiên cứu này sử dụng các độ đo tương tự để phân hoạch các đối tượng, nhưng chúng chỉ áp dụng cho các dữ liệu có thuộc tính số.
- Phân cụm khái niệm: Kỹ thuật này được phát triển áp dụng cho dữ liệu hạng mục, chúng phân cụm các đối tượng theo các khái niệm mà chúng xử lí.

- Phân cụm mờ: Sử đụng kỹ thuật mờ để phân cụm dữ liệu. Các thuật toán thuộc loại này chỉ ra lược đồ phân cụm thích hợp với tất cả các hoạt động đời sống hàng ngày, chúng chỉ xử lí các dữ liệu thực không chắc chắn.
- Phân cụm mạng Kohonen: Loại phân cụm này dựa trên khái niệm của các mạng nơron. Mạng Kohonen có tầng nơron vào và các tầng nơron ra. Mỗi nơron của tầng vào tương ứng với mỗi thuộc tính của bản ghi, mỗi một nơron vào kết nối với tất cả các nơron của tầng ra. Mỗi liên kết được gắn liền với một trọng số nhằm xác định vị trí của nơron ra tương ứng.



Hình 2.9: Các cách mà các cụm có thể đưa ra

2.5. Một số khái niệm cần thiết khi tiếp cận phân cụm dữ liệu

2.5.1. Phân loại các kiểu dữ liệu

Cho một CSDL *D* chứa n đối tượng trong không gian k chiều trong đó x, y, z là các đối tượng thuộc *D* : x =(x_1,x_2,..,x_k); y =(y_1,y_2,..,y_k); z =(z_1,z_2,..,z_k), trong đó x_i, y_i, z_i với i = 1…k là các đặc trưng hoặc thuộc tính tương ứng của các đối tượng x, y, z.

Sau đây là các kiểu dữ liệu:

a. Phân loại các kiểu dữ liệu dựa trên kích thước miền

- Thuộc tính liên tục (Continuous Attribute) : nếu miền giá trị của nó là vô hạn không đếm được
- Thuộc tính rời rạc (DiscretteAttribute) : Nếu miền giá trị của nó là tập hữu hạn, đếm được
- Lớp các thuộc tính nhị phân: là trường hợp đặc biệt của thuộc tính rời rạc mà miền giá trị của nó chỉ có 2 phần tử được diễn tả như : Yes / No hoặc Nam/Nữ, False/true,…

b. Phân loại các kiểu dữ liệu dựa trên hệ đo

Giả sử rằng chúng ta có hai đối tượng x, y và các thuộc tính x_i, y_i tương ứng với thuộc tính thứ i của chúng. Chúng ta có các lớp kiểu dữ liệu như sau :

- Thuộc tính định danh (Nominal Scale): đây là dạng thuộc tính khái quát hoá của thuộc tính nhị phân, trong đó miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử - nghĩa là nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là $x \neq y$ hoặc $x = y$.
- Thuộc tính có thứ tự (Ordinal Scale) : là thuộc tính định danh có thêm tính thứ tự, nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì ta có thể xác định là $x \neq y$ hoặc $x = y$ hoặc $x > y$ hoặc $x < y$.
- Thuộc tính khoảng (Interval Scale) : Với thuộc tính khoảng, chúng ta

có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu xi > yi thì ta nói x cách y một khoảng xi - yi tương ứng với thuộc tính thứ i.

- Thuộc tính tỉ lệ (Ratio Scale) : là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc, thí dụ như thuộc tính chiều cao hoặc cân nặng lấy điểm 0 làm mốc.Trong các thuộc tính dữ liệu trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục (Categorical), thuộc tính khoảng và thuộc tính tỉ lệ được gọi là thuộc tính số (Numeric).

2.5.2. Độ đo tương tự và phi tương tự

Để phân cụm, người ta phải đi tìm cách thích hợp để xác định "khoảng cách" giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính độ tương tự (Similar) hoặc là tính độ phi tương tự (Dissimilar) giữa các đối tượng dữ liệu.

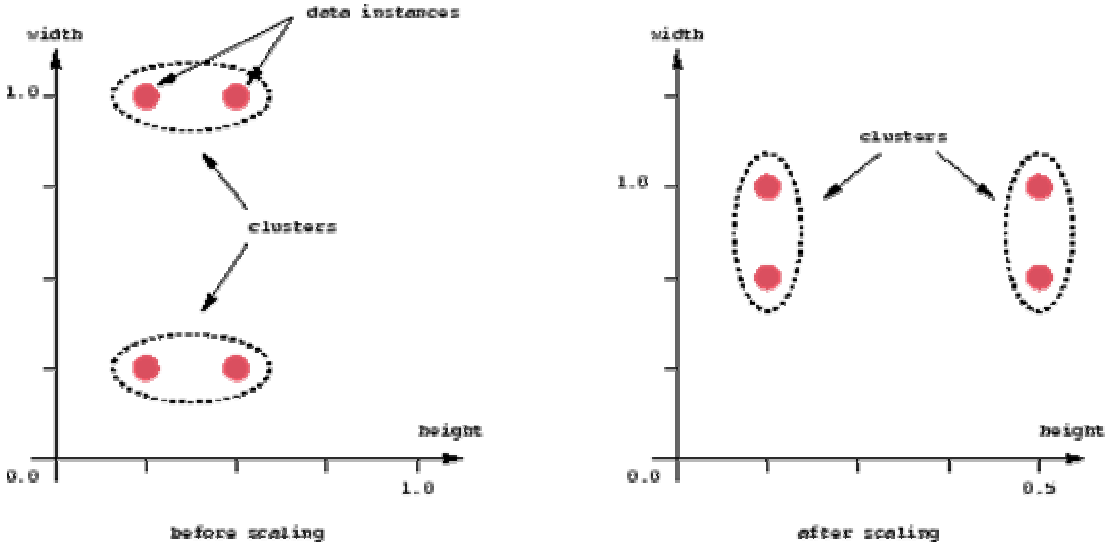
1. Không gian metric

Tất cả các độ đo dưới đây được xác định trong không gian độ đo metric. Một không gian metric là một tập trong đó có xác định các "khoảng cách" giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các phần tử của nó có thể là những đối tượng bất kỳ) các đối tượng dữ liệu trong CSDL D như đã đề cập ở trên được gọi là một không gian metric nếu:

- Với mỗi cặp phần tử x, y thuộc X đều có xác định, theo một quy tắc nào đó, một số thực $\delta(x,y)$, được gọi là khoảng cách giữa x và y.
- Quy tắc nói trên thoả mãn hệ tính chất sau : $\delta(x,y) > 0$ nếu $x \neq y$; (ii) $\delta(x, y)=0$ nếu x =y; (iii) $\delta(x,y) = \delta(y,x)$ với mọi x,y; (iv) $\delta(x,y) \leq$

δ(x,z)+δ(z,y).

Hàm δ(x,y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.



Hình 2.10: Minh họa số đo chiều rộng, chiều cao một đối tượng [8]
(phụ thuộc vào scaling khác nhau dẫn đến phân cụm khác nhau)

2. Thuộc tính khoảng cách:

Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các metric khoảng cách như sau [6, page 23]:

- Khoảng cách Minskowski: $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^q)^{1/q}$. trong đó q là số tự nhiên dương.
- Khoảng cách Euclide : $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$. đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=2.
- Khoảng cách Manhattan : $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$.

đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=1.

- Khoảng cách cực đại : $d(x,y) = Max_{i=1}^{n} \left| x_i - y_i \right|$.
đây là trường hợp của khoảng cách Minskowski trong trường hợp q→∞.

3. Thuộc tính nhị phân :

- α là tổng số các thuộc tính có giá trị là 1 trong x,y.
- β là tổng số các thuộc tính có giá trị là 1 trong x và 0 trong y.
- γ là tổng số các thuộc tính có giá trị là 0 trong x và 1 trong y.
- δ là tổng số các thuộc tính có giá trị là 0 trong x và y.
- $\tau = \alpha + \gamma + \beta + \delta$

Các phép đo độ tương tương đồng đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau :

Hệ số đối sánh đơn giản : $d(x,y) = \dfrac{\alpha + \delta}{\tau}$

ở đây cả hai đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

Hệ số Jacard : $d(x,y) = \dfrac{\alpha}{\alpha + \beta + \gamma}$

(bỏ qua số các đối sánh giữa 0-0). Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

4. Thuộc tính định danh :

Độ đo phi tương tự giữa hai đối tượng x và y được định nghĩa như sau:

$$d(x,y) = \frac{p-m}{p}$$

trong đó m là số thuộc tính đối sánh tương ứng trùng nhau, và p là tổng số các thuộc tính.

5. Thuộc tính có thứ tự :

Giả sử i là thuộc tính thứ tự có Mi giá trị (M_i kích thước miền giá trị) : Các trạng thái Mi được sắp thứ tự như sau : [1…M_i], chúng ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại r_i, với $r_i \in \{1 \ldots M_i\}$.

Mỗi một thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy chúng ta chuyển đổi chúng về cùng miền giá trị [0,1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính :

$$Z_i^{(j)} = \frac{r_i^{(j)} - 1}{M_i - 1}$$

Sử dụng công thức tính độ phi tương tự của thuộc tính khoảng đối với các giá trị $Z_i^{(j)}$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

6. Thuộc tính tỉ lệ :

Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính. Hoặc loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng w_i (1 <= i <= k), độ tương đồng dữ liệu được xác định như sau :

$$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$$

2.6. Một số thuật toán cơ bản trong phân cụm dữ liệu

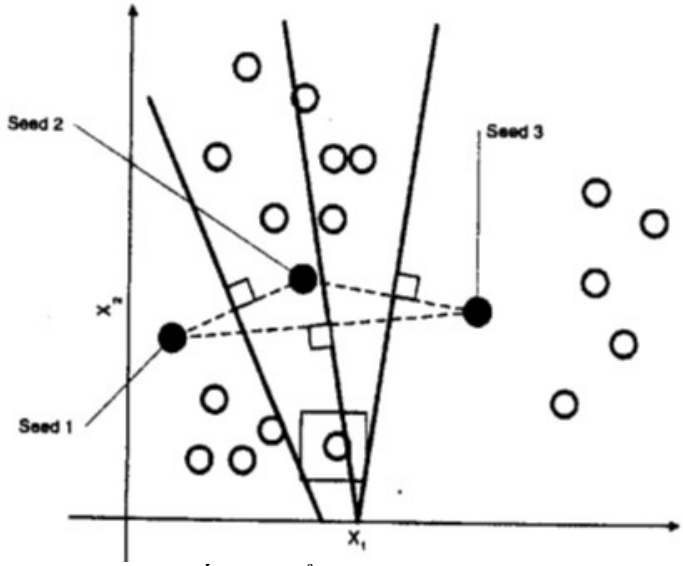
2.6.1. Các thuật toán phân cụm phân hoạch

Cho trước một cơ sở dữ liệu với n đối tượng hay các bộ dữ liệu, một phương pháp phân chia được xây dựng để chia dữ liệu thành k phần, mỗi phần đại diện cho một cụm $k \leq n$. Đó là phân loại dữ liệu vào trong k nhóm, chúng thoả các yêu cầu sau : (1) Mỗi nhóm phải chứa ít nhất một đối tượng; (2) Mỗi đối tượng phải thuộc về chính xác một nhóm. (yêu cầu thứ 2 được nới lỏng trong kỹ thuật phân chia cụm mờ).

Có rất nhiều thuật toán phân hoạch như : k-means (MacQueen 1967), k-medoids (Kaufman và Rousseew 1987), PAM (Partition Around Medoids), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based on RAndomized Search), CLASA (Clustering Large Applications based on Simulated Annealing).

1. Thuật toán k-mean [7]

Thuật toán này dựa trên độ đo khoảng cách của các đối tượng dữ liệu đến phần tử là trung tâm của cụm chứa nó.



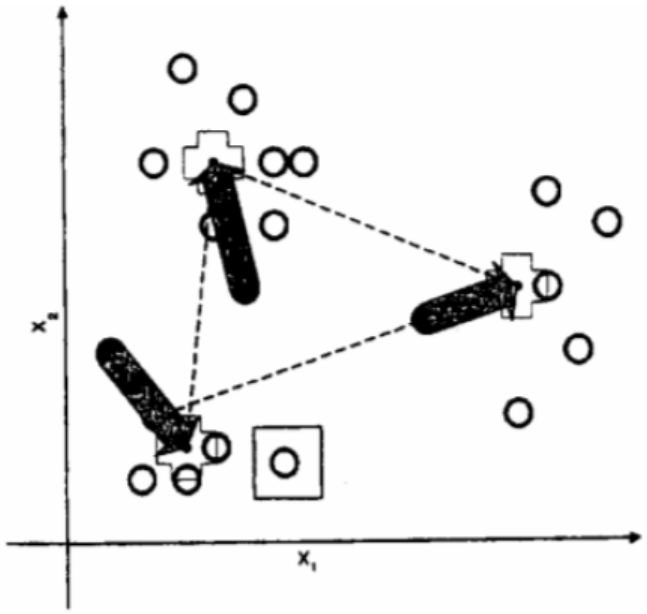
Hình 2.11: Các thiết lập để xác định ranh giới các cụm ban đầu

Thuật toán k-means lấy tham số đầu vào là k và phân chia một tập n đối tượng vào trong k cụm để cho kết quả độ tương đồng trong cụm là cao trong khi độ tương đồng ngoài cụm là thấp. Độ tương đồng cụm được đo khi đánh giá giá trị trung bình của các đối tượng trong cụm, nó có thể được quan sát như là "trọng tâm" của cụm.

Giải thuật xử lý như sau: trước tiên nó lựa chọn ngẫu nhiên k đối tượng, mỗi đối tượng đại diện cho một trung bình cụm hay tâm cụm. Đối với những đối tượng còn lại, mỗi đối tượng sẽ được ấn định vào một cụm mà nó giống nhất dựa trên khoảng cách giữa đối tượng và trung bình cụm. Sau đó sẽ tính lại trung bình cụm mới cho mỗi cụm. Xử lý này sẽ được lặp lại cho tới khi hàm tiêu chuẩn hội tụ. Bình phương sai số thường dùng làm hàm tiêu chuẩn hội tụ, định nghĩa như sau :

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$$

với x là điểm trong không gian đại diện cho đối tượng cho trước, m_i là trung bình cụm C_i (cà x và m_i đều là đa chiều). Tiêu chuẩn này cố gắng cho kết quả k cụm càng đặc, càng riêng biệt càng tốt.



Hình 2.12: Tính toán trọng tâm của các cụm mới

Thuật toán k-means bao gồm các bước cơ bản sau :

Đầu vào : Số cụm k và hàm E $E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$

Đầu ra : Các cụm C[i] ($1 \le i \le k$) với hàm tiêu chuẩn E đạt giá trị tối thiểu.

Begin

Bước 1 : Khởi tạo

Chọn ngẫu nhiên k tâm $\{m_j\}_{j=1}^{k}$ ban đầu trong không gian R^d (d là số chiều của dữ liệu). Mỗi cụm được đại diện bằng các tâm của cụm .

Bước 2: Tính toán khoảng cách $D_{j=1}^{k} \sqrt{\sum_{i=1}^{n} (x_i - m_j)^2}$

Đối với mỗi điểm x_i($1 \le i \le n$), tính toán khoảng cách của nó tới mỗi trọng tâm m_j($1 \le j \le k$). Sau đó tìm trọng tâm gần nhất đối với mỗi điểm và nhóm chúng vào các nhóm gần nhất.

Bước 3: Cập nhật lại trọng tâm

Đối với mỗi $1 \le j \le k$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu.

Bước 4: Gán lại các điểm gần trung tâm nhóm mới

Nhóm các đối tượng vào nhóm gần nhất dựa trên trọng tâm của nhóm.

Điều kiện dừng:

Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

End.

Thuật toán k-means được chứng minh là hội tụ và có độ phức tạp tính

toán là O(tkn) với t là số lần lặp, k là số cụm, n là số đối tượng của tập dữ liệu vào. Thông thường k<<n và t<<n thường kết thúc tại một điểm tối ưu cục bộ.

Tuy nhiên, nhược điểm của k-means là còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu. Hơn nữa, chất lượng phân cụm dữ liệu của thuật toán k-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k trọng tâm khởi tạo ban đầu. Trong trường hợp các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của k-means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào, giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào k khác nhau rồi sau đó chọn giải pháp tốt nhất.

Đánh giá thuật toán K-Means

- Ưu điểm :

- K-means là có độ phức tạp tính toán O(tkn).
- K-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.

- Nhược điểm :

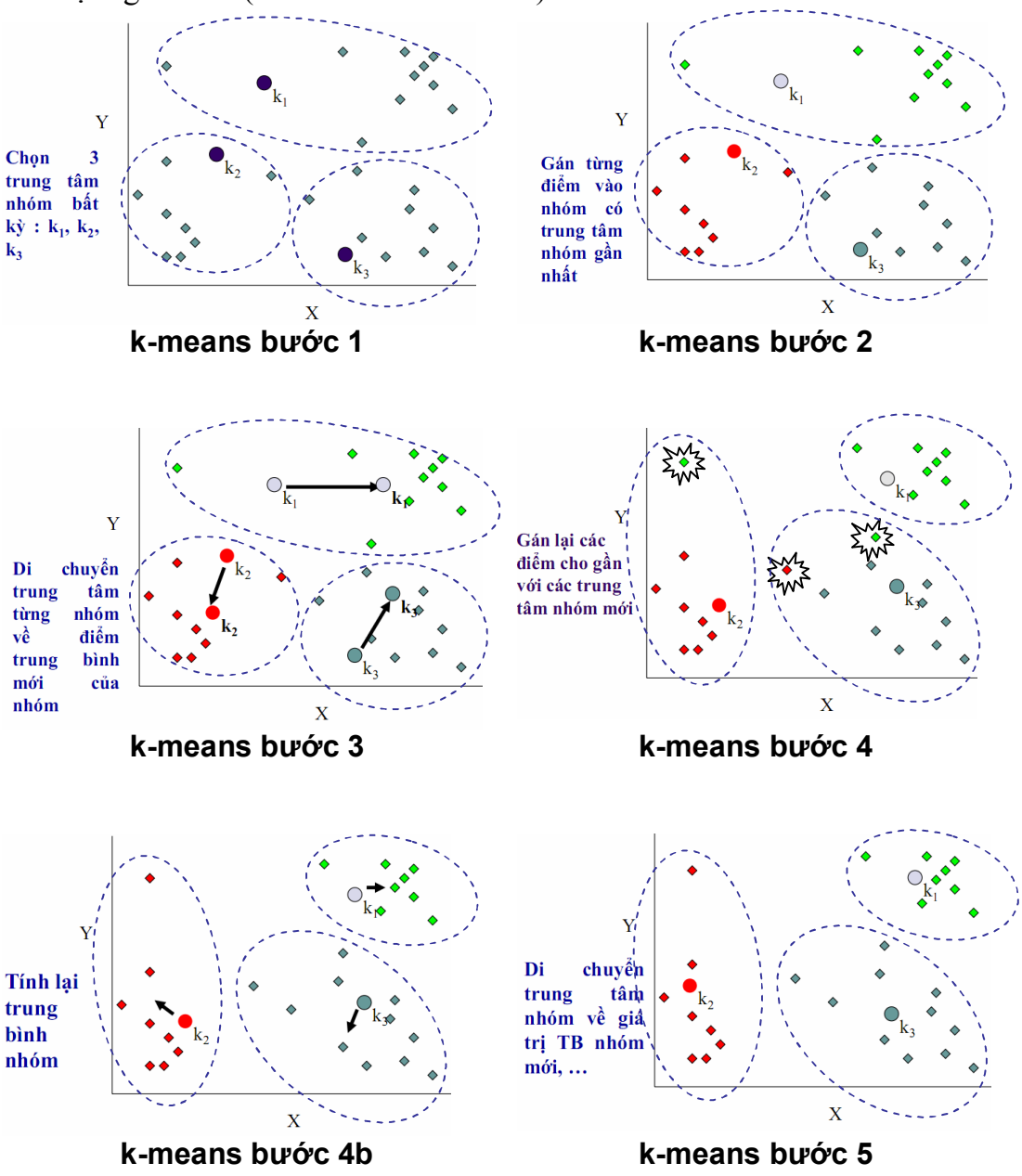
- K-means không khắc phục được nhiễu và giá trị k phải được cho bởi người dùng.
- Chỉ thích hợp áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu.

Ví dụ : Giả sử có một tập đối tượng được định vị trong hệ trục toạ độ X, Y. Cho k =3 tức người dùng cần phân các đối tượng vào trong 3 cụm.

Theo giải thuật, ta chọn ngẫu nhiên 3 trung tâm cụm ban đầu (Hình k-means bước 1). Sau đó, mỗi đối tượng được phân vào trong các cụm đã chọn

dựa trên tâm cụm gần nhất (Hình k-means bước 2).

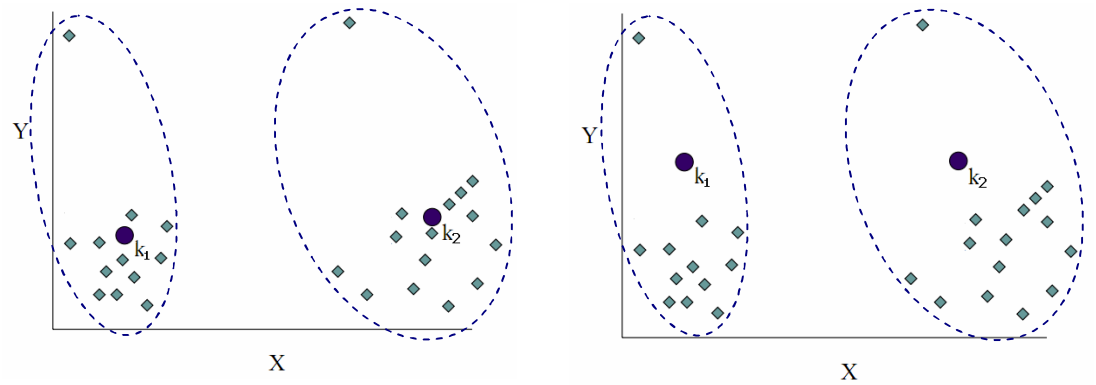
Cập nhật lại các tâm (Hình k-means bước 3). Đó là giá trị trung bình của mỗi cụm được tính toán lại dựa trên các đối tượng trong cụm. Tuỳ theo các tâm mới này, các đối tượng được phân bố lại vào trong các cụm dựa trên tâm cụm gần nhất (Hình k-means bước 4).



Hình 2.13: Ví dụ các bước của thuật toán k-means

2. Thuật toán PAM

Giải thuật k-means rất nhạy với các phần tử ngoại lai, do vậy một đối tượng giá trị cực lớn về cơ bản sẽ làm thay đổi tâm cụm và có thể bóp méo phân bổ của dữ liệu.



Tâm cụm không có phần tử ngoại lai **Tâm cụm khi có phần tử ngoại lai**

Hình 2.14: Sự thay đổi tâm cụm trong k-means khi có phần tử ngoại lai

Ý tưởng của k-medodis thay vì lấy giá trị trung bình của các đối tượng trong cụm như một điểm tham khảo, k-medoids lấy một đối tượng đại diện trong cụm, gọi là medoid, nó là điểm đại diện được định vị trung tâm nhất trong cụm. Do vậy, phương pháp phân chia vẫn được thực hiện dựa trên nguyên tắc tối thiểu hoá tổng các độ không tương động giữa mỗi đối tượng với điểm tham khảo tương ứng của nó, điểm này thiết lập nên cơ sở của phương pháp k-mediods.

Giải thuật PAM, đây là giải thuật phân cụm kiểu k-mediods. Nó tìm k cụm trong n đối tượng bằng cách trước tiên tìm một số đối tượng đại diện (medoid) cho mỗi cụm. Tập các medoid ban đầu được lựa chọn tuỳ ý. Sau đó nó lặp lại các thay một trong số các medoid bằng một trong số những cái không phải medoid miễn là tổng khoảng cách của kết quả phân cụm được cải thiện.

Giải thuật thử xác định k phần phân chia cho n đối tượng. sau khi lựa

chọn được k-medoids ban đầu, giải thuật lặp lại việc thử để có một sự lựa chọn các medoid tốt hơn bằng cách phân tích tất cả các cặp đối tượng có thể để một đối tượng là medoid và đối tượng kia thì không phải. Phép đo chất lượng phân cụm được tính cho mỗi sự kết hợp như vậy. Lựa chọn các điểm tốt nhất trong một lần lặp được chọn với tư cách là các medoid cho lần lặp tiếp theo. Độ phức tạp cho một lần lặp đơn là $O(k(n - k)^2)$, với độ phức tạp như trên không thich hợp cho phân cụm dữ liệu có số lượng n lớn và số cụm cần chia là nhiều.

Thuật toán PAM bao gồm các bước cơ bản sau :

Đầu vào : Số cụm k và một cơ sở dữ liệu chứa n đối tượng

Đầu ra : Một tập k cụm đã tối thiểu hoá tổng các độ đo không tương đồng của tất cả các đối tượng tới medoid gần nhất của chúng

Bắt đầu

1. Chọn tuỳ ý k đối tượng giữ vai trò là các medoid ban đầu;
2. Repeat
3. Ấn định mỗi đối tượng vào cụm có medoid gần nó nhất;
4. Tính hàm mục tiêu (tổng các độ đo tương đồng của tất cả các đối tượng tới medoid gần nhất cùa chúng);
5. Đổi medoid x bằng một đối tượng y nếu như việc thay đổi này làm giảm hàm mục tiêu;
6. Until : không có sự thay đổi nào

Kết thúc

Khi có sự hiện diện của nhiễu và các phần tử ngoại lai, phương pháp m-medoids mạnh hơn k-means bởi so với giá trị trung bình (mean), medoid ít

bị ảnh hưởng hơn bởi các phần tử ngoại lai hay các giá trị ở rất xa khác nữa. Tuy nhiên, xử lý nó tốn thời gian hơn so với k-means

3. Thuật toán CLARA

Thuật toán PAM làm việc hiệu quả đối với các tập dữ liệu nhỏ nhưng không có khả năng mở rộng tốt đối với các tập dữ liệu lớn, trong trường hợp giá trị k và n là lớn. Để giải quyết các dữ liệu lớn, một phương pháp dựa trên việc lấy mẫu gọi là CLARA (Clustering large applications) được phát triển bởi Kaufman va Rousseeuw năm 1990.

Ý tưởng của CLARA như sau : thay vì lấy toàn bộ dữ liệu vào xem xét, chỉ một phần nhỏ dữ liệu được chọn với vai trò là một đại diện của dữ liệu, và các medoid được chọn từ mẫu này bằng cách sử dụng PAM. Nếu như mẫu được chọn lựa khá ngẫu nhiên, nó đại diện phù hợp cho toàn bộ tập dữ liệu và các đối tượng đại diện (các medoid) được chọn do vậy sẽ giống với những cái được chọn lựa từ toàn bộ tập dữ liệu. CLARA đưa ra nhiều mẫu của tập dữ liệu, áp dụng PAM trên từng mẫu và mang lại phân cụm tốt cho đầu ra. Đúng như trông chờ, CLARA có thể giải quyết với các tập dữ liệu lớn hơn PAM. Độ phức tạp của mỗi lần lặp bây giờ trở thành $O(kS^2+k(n – k))$ với S là kích thước mẫu, k là số cụm, n là tổng số các phần tử.

Hiệu quả của CLARA tuỳ thuộc vào kích thước mẫu. Lưu ý rằng PAM tìm kiếm cho k medoids tốt nhất giữa một tập dữ liệu cho trước, trong khi đó CLARA tìm kiếm cho k medoids tốt nhất giữa các mẫu đã lựa chọn của tập dữ liệu. CLARA không thể tìm được phân cụm tốt nhất nếu như bất kỳ một medoid được lấy mẫu không nằm trong k medoids tốt nhất. Ví dụ, nếu một đối tượng O_i là một trong k medoids tốt nhất nhưng nó không được chọn trong suốt quá trình lấy mẫu, CLARA sẽ không bao giờ tìm thấy phân cụm tốt nhất. Một phân cụm tốt dựa trên các mẫu chưa chắc đã địa diện cho một phân cụm tốt cho toàn bộ dữ liệu nếu mẫu bị lệch (bias).

4. Thuật toán CLARANS

Để cải thiện chất lượng và khả năng mở rộng của CLARA, một giải thuật phân cụm khác gọi là CLARANS (Clustering Large Applications based upon RANdomized Search) giới thiệu bởi Ng và Han năm 1994. Nó cũng là một giải thuật kiểu k-medoids và kết hợp kỹ thuật lấy mẫu với PAM. Tuy vậy, không giống như CLARA, CLARANS không hạn chế bản thân nó cho bất kỳ một mẫu nào tại bất kỳ thời điểm nào cho trước. Trong khi đó CLARA lại có một mẫu được ấn định tại mọi giai đoạn tìm kiếm, CLARANS đưa ra một mẫu một cách ngẫu nhiên trong mỗi bước tìm kiếm. Xử lý phân cụm được thực hiện như tìm kiếm một đồ thị tại mọi nút là giải pháp tiềm năng, tức là một tập k-medoids. Phân cụm có được sau khi thay thế một medoid được gọi là láng giềng của phân cụm hiện thời. Số lượng các láng giềng được thử ngẫu nhiên bị hạn chế bởi một tham số. Nếu như một láng giềng tốt hơn được tìm thấy CLARANS di chuyển tới láng giềng đó và bắt đầu xử lý lại; ngược lại, phân cụm hiện thời đưa ra một tối ưu cục bộ. Nếu như tối ưu cục bộ được tìm thấy, CLARANS bắt đầu với các nút được lựa chọn ngẫu nhiên mới để tìm kiếm một tối ưu cục bộ mới. Bằng thực nghiệm, CLARANS đã chỉ ra là hiệu quả hơn PAM và CLARA. Độ phức tạp tính toán của mỗi lần lặp trong CLARANS tỷ lệ tuyến tính với số lượng các đối tượng. CLARANS được dùng để tìm số lượng lớn nhất các cụm tự nhiên sử dụng hệ số hình chiếu- đây là một đặc tính của các phần tử ngoại lai, tức là các điểm mà không thuốc bất kỳ cụm nào.

Một số khái niệm sử dụng trong thuật toán CLARANS được định nghĩa như sau:

Giả sử O là một tập có n đối tượng và $M \subseteq O$ là tập các đối tượng tâm mediod, NM = O - M là tập các đối tượng không phải tâm. Các đối tượng dữ liệu sử dụng trong thuật toán CLARANS là các khối đa diện. Mỗi đối tượng được diễn tả bằng một tập các cạnh, mỗi cạnh được xác định bằng hai điểm.

Giả sử $P \subseteq R^3$ là một tập tất cả các điểm. Nói chung, các đối tượng ở đây là các đối tượng dữ liệu không gian và chúng ta định nghĩa tâm của một đối tượng chính là trung bình cộng toán học của tất cả các đỉnh hay còn gọi là trọng tâm :

center : $O \rightarrow P$

Giả sử dist là một hàm khoảng cách, khoảng cách thường được chọn ở đây là khoảng cách Eucliean : dist : PxP $\rightarrow R_0^+$

Hàm khoảng cách dist có thể mở rộng cho các điểm của khối đa diện thông qua hàm tâm : dist : OxO $\rightarrow R_0^+$ sao cho

$dist(o_i,o_j) = dist(center(o_i),center(o_j))$

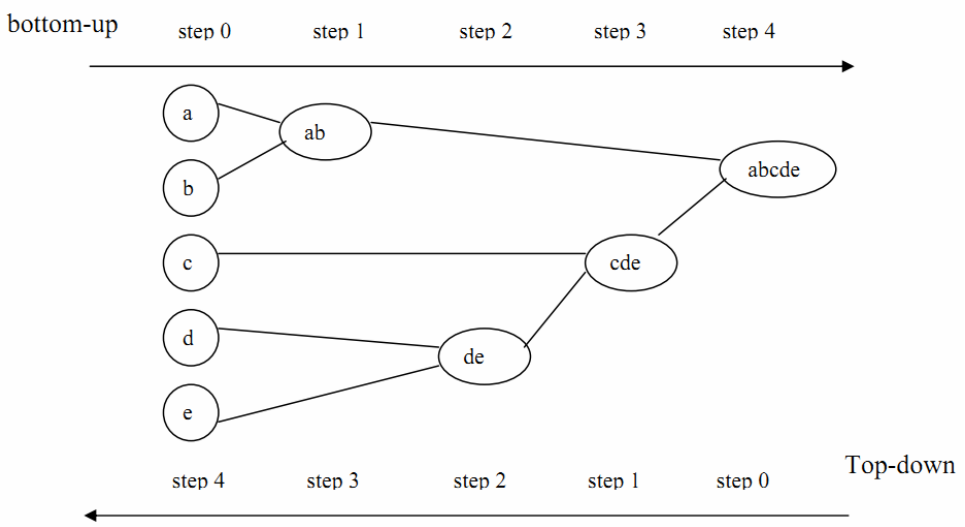
Mỗi đối tượng được gán cho một tâm medoid của cụm nếu khoảng cách từ trọng tâm của đối tượng đó tới tâm medoid của nó là nhỏ nhất. Vì vậy, định nghĩa tâm medoid như sau : medoid : O $\rightarrow$ M sao cho :

$medoid(o) = m_i$, $m_i \in M, \forall m_i \in M : dis(o,m_i)$ $dist(o,m_j)$, $o \in O$. Cuối cùng định nghĩa một cụm tới tâm mediod m_i tương ứng là một tập con các đối tượng trong O với $medoid(o) = m_i$

Giả sử C0 là tập tất cả các phân hoạch của O. Hàm tổng để đánh giá chất lượng một phân hoạch được định nghĩa như sau: total_distance:$C_0 \rightarrow R_0^+$ sao cho $total\_distance(c) = \sum\sum dist(o,m_i)$ với $m_i \in M$, $o \in cluster(m_i)$

2.6.2. Các thuật toán phân cụm phân cấp

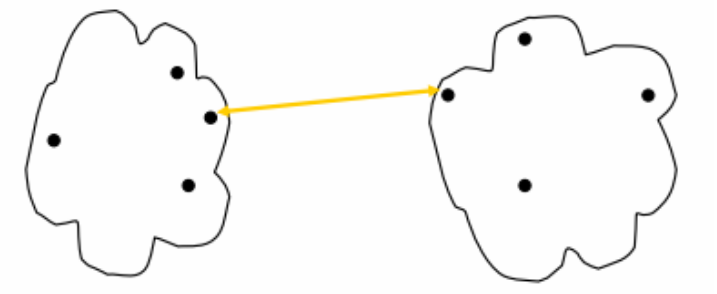
Phương pháp phân cụm phân cấp làm việc bằng cách nhóm các đối tượng dữ liệu vào trong một cây các cụm.



Hình 2.15 : Phân cụm phân cấp Top-down và Bottom-up

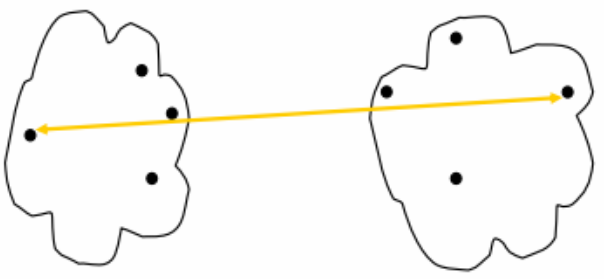
Trong phương pháp phân cụm phân cấp cần nhắc lại cách xác định khoảng cách giữa 2 nhóm [6, page 36]:

- Single Link : khoảng cách ngắn nhất giữa hai đối tượng thuộc hai nhóm



Hình 2.16 : Single Link

- Complete Link : khoảng cách xa nhất giữa hai đối tượng thuộc hai nhóm



Hình 2.17 : Complete Link

Các thuật toán điển hình của phương pháp phân cụm phân cấp đó là: ANGNES (Agglomerative Nesting), DIANA (Divisive Analysis), BIRCH (Balanced Interative Reducing and Clustering using Hierarchies), CURE (Clustering Using REpresentatives), ROCK, Chameleon ...

2. Thuật toán AGNES

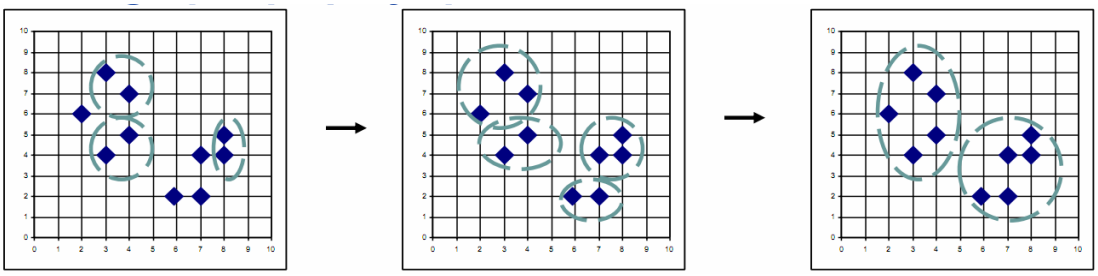
Phương pháp phân cụm AGNES là kỹ thuật kiểu tích tụ. AGNES bắt đầu ở ngoài với mỗi đối tượng dữ liệu trong các cụm riêng lẻ. Các cụm được hòa nhập theo một số loại của cơ sở luật, cho đến khi chỉ có một cụm ở đỉnh của phân cấp, hoặc gặp điều kiện dừng. Hình dạng này của phân cụm phân cấp cũng liên quan đến tiếp cận bottom-up bắt đầu ở dưới với các nút lá trong mỗi cụm riêng lẻ và duyệt lên trên phân cấp tới nút gốc, nơi tìm thấy cụm đơn cuối cùng với tất cả các đối tượng dữ liệu được chứa trong cụm đó.

Thuật toán AGNES bao gồm các bước cơ bản sau :

Bước 1: Mỗi đối tượng là một nhóm

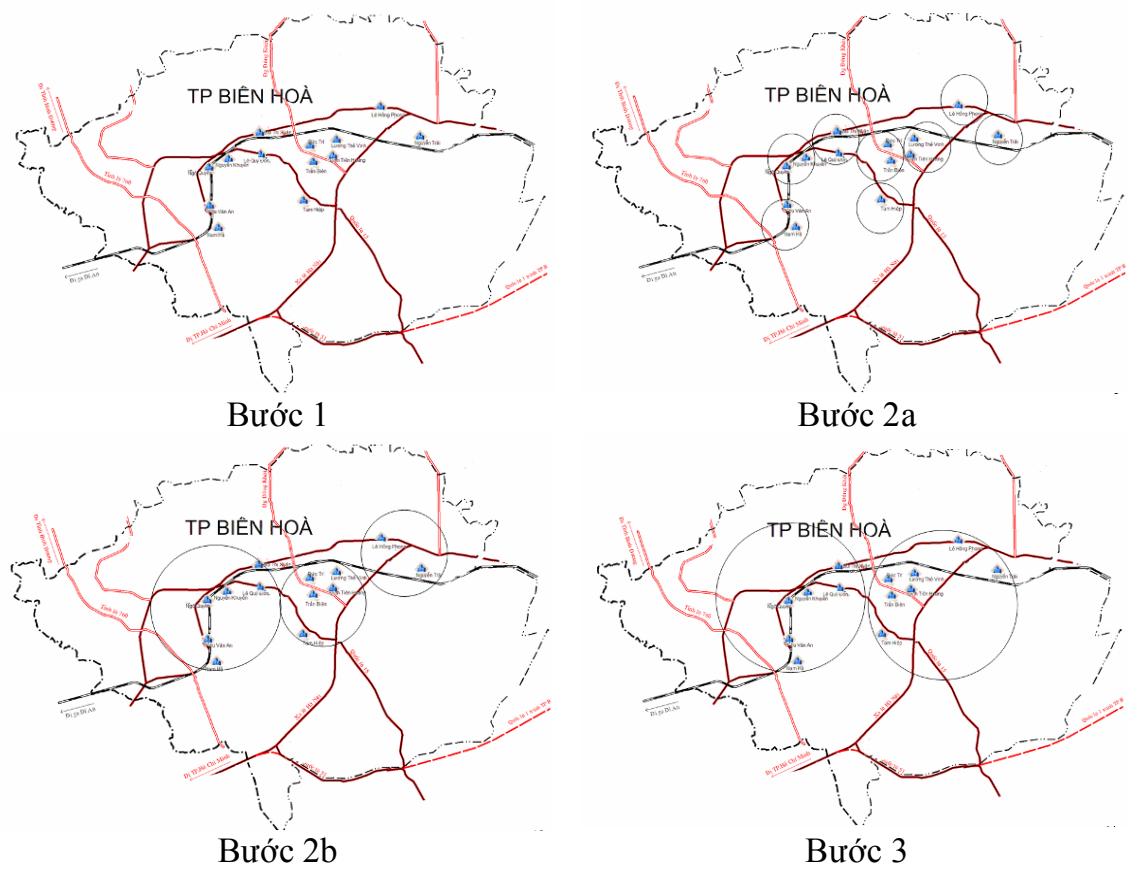
Bước 2: Hợp nhất các nhóm có khoảng cách giữa các nhóm là nhỏ nhất (Single Link).

Bước 3: Nếu thu được nhóm "toàn bộ" thì dừng, ngược lại quay lại bước 2.



Hình 2.18 : Các bước cơ bản của AGNES [7]

Ví dụ : Sử dụng thuật toán AGNES để phân cụm thi các trường trong nội ô thành phố Biên Hoà thông qua Single Link



Hình 2.19: Ví dụ các bước cơ bản của thuật toán AGNES

3. Thuật toán DIANA

DIANA thực hiện đối lập với AGNES. DIANA bắt đầu với tất cả các đối tượng dữ liệu được chứa trong một cụm lớn và chia tách lặp lại, theo phân loại giống nhau dựa trên luật, cho đến khi mỗi đối tượng dữ liệu của cụm lớn được chia tách hết. Hình dạng của cụm phân cấp cùng liên quan để tiếp cận top-down bắt đầu tại mức đỉnh nút gốc, với tất cả các đối tượng dữ liệu, trong một cụm, và duyệt xuống các nút lá dưới cùng nơi tất cả các đối tượng dữ liệu từng cái được chứa trong cụm của chính mình.

Trong mỗi phương pháp của hai phương pháp, có thể số các cụm dẫn tới các mức khác nhau trong phân cấp bằng cách duyệt lên hoặc xuống cây. Mỗi mức có thể khác nhau số các cụm và tất nhiên kết quả cũng khác nhau.

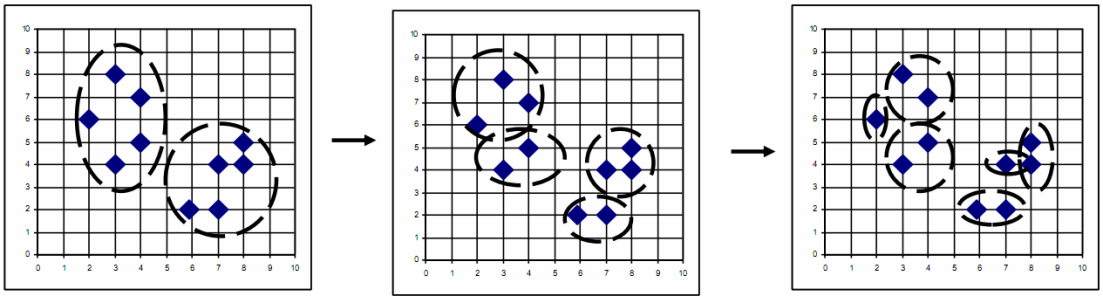
Một hạn chế lớn của cách tiếp cận này là các cụm được hòa nhập hoặc phân chia một lần, không thể quay lại quyết định đó, cho dù hòa nhập hoặc phân chia không phải là thích hợp ở mức đó

Thuật toán DIANA bao gồm các bước cơ bản sau :

Bước 1: Tất cả các đối tượng là một nhóm

Bước 2: Chia nhỏ nhóm có khoảng cách giữa những đối tượng trong nhóm là lớn nhất (Complete Link).

Bước 3: Nếu mỗi nhóm chỉ chứa một đối tượng thì dừng, ngược lại quay lại quay lại bước 2.

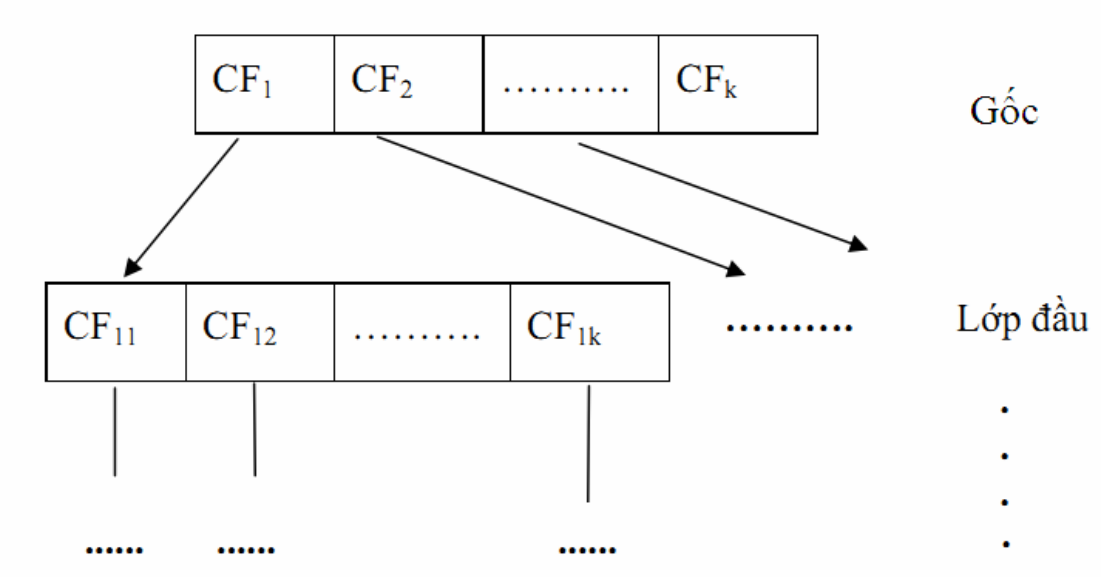


Hình 2.20 : Các bước cơ bản của DIANA [7]

Cả 2 thuật toán AGNES và DIANA về cơ bản mặc dù đơn giản nhưng thường gặp khó khăn khi ra các quyết định tới hạn cho việc lựa chọn của điểm hoà nhập hay phân chia một cách chính xác. Quyết định như vậy gọi là tới hạn bởi một khi một nhóm các đối tượng được hoà nhập hay chia, xử lý tại bước tiếp theo sẽ làm việc trên trên các cụm mới sinh ra. Nó sẽ không bao giờ huỷ những việc đã làm trước đó và cũng không thực hiện chuyển đổi đối tượng giữa các cụm. Do vậy các quyết định hoà nhập hay phân chia nếu không đủ sáng suốt ở mỗi bước thì có thể dẫn tới chất lượng các cụm sẽ kém. Hơn nữa, phương pháp này khả năng mở rộng không được tốt nên quyết định hoà nhập hay phân chia cần kiểm định và đánh giá một số lượng tốt các đối tượng hay các cụm.

4. Thuật toán BIRCH

BIRCH là thuật toán phân cụm phân cấp sử dụng chiến lược Top-down. Tư tưởng của BIRCH là không lưu toàn bộ đối tượng dữ liệu của các cụm trong bộ nhớ mà chỉ lưu các tham số thống kê. Đối với mỗi cụm dữ liệu, BIRCH chỉ lưu bộ ba (N, LS, SS), trong đó N là số đối tượng trong cụm, LS là tổng các giá trị thuộc tính của các đối tượng trong cụm, và SS là tổng bình phương của các giá trị thuộc tính của các đối tượng trong cụm. Bộ ba này được gọi là đặc trưng cụm (Cluster Feature - CF). Khi đó các cụm trong tập dữ liệu ban đầu sẽ được cho dưới dạng một cây CF. Người ta đã chứng minh được rằng các đại lượng thống kê như độ đo có thể xác định từ cây CF.



Hinh 2.21 : Cấu trúc cây CF

Cây CF là một cây cân bằng nhằm lưu các đặc trưng của cụm. Một cây CF chứa các nút cha và lá, nút cha chứa các nút con, nút lá không có con. Nút cha lưu giữ tổng các đặc trưng cụm của các nút con của nó. Cây CF có hai đặc trưng cơ bản sau:

a. Yếu tố nhánh (Branching Factor- B) nhằm xác định số lượng nút con tối đa trong một nút cha.

b. Ngưỡng (Threshold- T) nhằm xác định khoảng cách tối đa giữa bất kỳ một cặp đối tượng trong nút lá của cây CF, khoảng cách này chính là đường kính của các cụm con được lưu lại ở nút lá.

Thuật toán BIRCH được thực hiện qua hai giai đoạn sau:

Giai đoạn 1 : Duyệt tất cả các đối tượng trong tập dữ liệu và xây dựng một cây CF ban đầu. Ở giai đoạn này các đối tượng lần lượt được chèn vào nút lá gần nhất của cây CF (nút lá của cây đóng vai trò cụm con), sau khi chèn xong thì mọi nút trên cây CF được cập nhật thông tin. Nếu đường kính của cụm con sau khi chèn lớn hơn ngưỡng T thì nút được tách. Quá trình này được lặp đi lặp lại cho đến khi tất cả các đối tượng đều được chèn vào cây CF.

Giai đoạn 2 : BIRCH chọn một giải thuật toán phân cụm bất kỳ (như thuật toán phân hoạch) để thực hiện phân cụm cho tất các các nút lá CF.

Đánh giá thuật toán BIRCH.

- Ưu điểm:

Nhờ sử dụng cây CF, BIRCH có tốc độ phân cụm nhanh độ phức tạp O(n) (vì BIRCH chỉ duyệt toàn bộ dữ liệu một lần). BIRCH được áp dụng đối với tập dữ liệu lớn, đặc biệt phù hợp với các dữ liệu gia tăng theo thời gian.

- Nhược điểm:

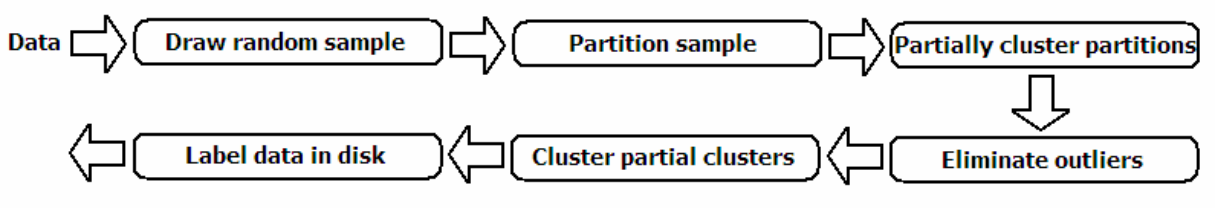
Chất lượng cụm được khám phá bởi BIRCH là không tốt. Tham số ngưỡng T ảnh hưởng lớn đến kích thước và tính tự nhiên của cụm.

5. Thuật toán CURE

Trong khi hầu hết các thuật toán thực hiện phân cụm với các cụm hình cầu và kích thước tương tự, như vậy là không hiệu quả khi xuất hiện các phần tử ngoại lai. Thuật toán CURE khắc phục được vấn đề này và tốt hơn với các

phần tử ngoại lai.

CURE là thuật toán sử dụng chiến lược bottom-up của phương pháp phân cụm phân cấp. Khác với các thuật toán phân cụm phân hoạch, thuật toán CURE sử dụng nhiều đối tượng để biểu diễn cho một cụm thay vì sử dụng các trọng tâm hay đối tượng tâm. Các đối tượng đại diện của một cụm ban đầu được chọn rải rác đều ở các vị trí khác nhau, sau đó chúng được di chuyển bằng cách co lại theo một tỉ lệ nhất định nào đó, quá trình này được lặp lại và nhờ vậy trong quá trình này, có thể đo tỉ lệ gia tăng của cụm. Tại mỗi bước của thuật toán, hai cụm có cặp các điểm đại diện gần nhau (mỗi điểm trong cặp thuộc về mỗi cụm khác nhau) được hòa nhập hai đối tượng đại diện gần nhất sẽ được trộn lại thành một cụm.



Hình 2.22 : Khái quát thuật toán CURE

Để xử lí được các CSDL lớn, CURE sử dụng mẫu ngẫu nhiên và phân hoạch, một mẫu là được xác định ngẫu nhiên trước khi được phân hoạch và sau đó tiến hành phân cụm trên mỗi phân hoạch, như vậy mỗi phân hoạch là từng phần đã được phân cụm, các cụm thu được lại được phân cụm lần thứ hai để thu được các cụm con mong muốn, nhưng mẫu ngẫu nhiên không nhất thiết đưa ra một mô tả tốt cho toàn bộ tập dữ liệu.

Các bước thực hiện của thuật toán CURE:

a. Chọn một mẫu ngẫu nhiên S từ tập dữ liệu ban đầu.

b. Phân hoạch mẫu S thành các nhóm dữ liệu có kích thước bằng nhau.

c. Tiến hành phân cụm riêng rẽ cho mỗi nhóm.

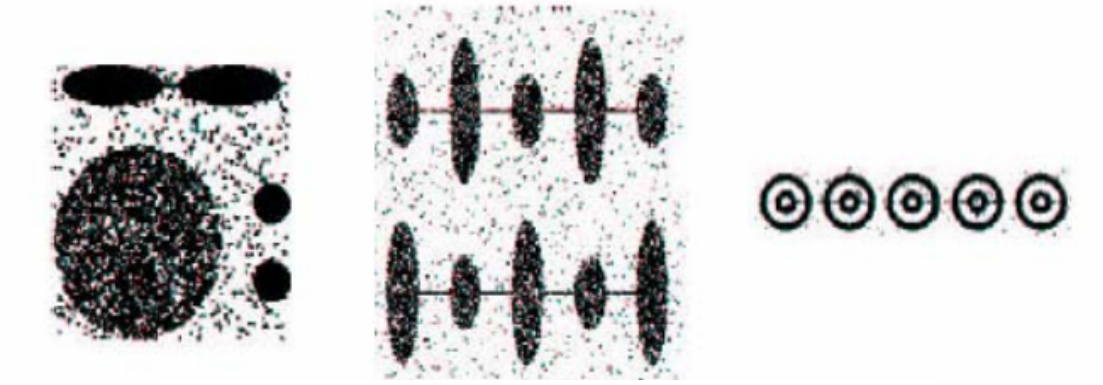
d. Loại bỏ các đối tượng ngoại lai bằng việc lấy mẫu ngẫu nhiên. Nếu một cụm tăng trưởng quá chậm thì loại bỏ nó.

e. Phân cụm cho các cụm riêng biệt: Các đối tượng đại diện được di chuyển về phía tâm của cụm mới hình thành. Các đối tượng này sẽ mô tả hình dạng cụm đó.

f. Đánh dấu dữ liệu với các nhãn cụm tương ứng.

Độ phức tạp tính toán của thuật toán CURE là $O(n^2 1og(n))$. CURE là thuật toán tin cậy trong việc khám phá ra các cụm với hình thù bất kỳ và có thể áp dụng tốt đối với dữ liệu có phần tử ngoại lai và trên các tập dữ liệu hai chiều. Tuy nhiên, nó lại rất nhạy cảm với các tham số như số các đối tượng đại diện, tỉ lệ co của các phần tử đại điện.

Hình ảnh dưới đây là thí dụ về các dạng và kích thước cụm dữ liệu được khám phá bởi CURE :



Hinh 2.23 : Các cụm dữ liệu được khám phá bởi CURE

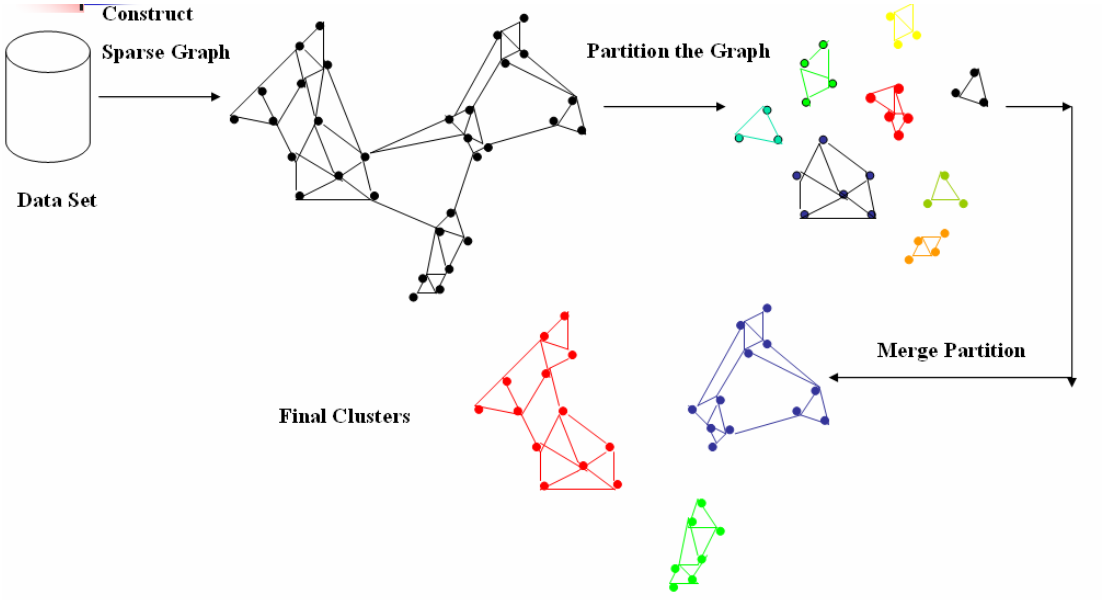
6. Thuật toán Chameleon

Phương pháp Chameleon một cách tiếp cận khác trong việc phân cụm được phát triển bởi Karypis, Han và Kumar năm 1999, sử dụng mô hình động trong phân cụm phân cấp.

Khi xử lý phân cụm, 2 cụm được hoà nhập nếu liên kết nối và độ chặt (độ gần) giữa hai cụm được liên kết cao với liên kết nối và độ chặt nội tại của các đối tượng nằm trong phạm vi các cụm. Xử lý hoà nhập dựa trên mô hình động tạo điều kiện thuận lợi cho khám phá ra các cụm tự nhiên và đồng nhất, nó áp dụng cho tất cả các kiểu dữ liệu miễn là hàm tương đồng được chỉ định.

CHAMELEON có được dựa trên quan sát các yếu điểm của giải thuật phân cụm phân cấp CURE, ở đó CURE và các lược đồ đã bỏ qua thông tin về liên kết của các đối tượng trong hai cụm khác nhau.

Trước đầu tiên của Chameleon là xây dựng một đồ thị mật độ thưa và sau đó ứng dụng một thuật toán phân hoạch đồ thị để phân cụm dữ liệu với số lớn của các cụm con. Tiếp theo, Chameleon thực hiện tích tụ phân cụm phân cấp như AGNES, bằng hòa nhập các cụm con nhỏ theo hai phép đo, mối quan hệ liên thông và mối quan hệ gần nhau của các nhóm con. Do đó, thuật toán không phụ thuộc vào người sử dụng các tham số như K-means.



Hình 2.24 : Khái quát thuật toán CHAMELEON [7]

Như vậy, nó không phụ thuộc vào mô hình tĩnh hay động và có thể từ động thích nghi với đặc trưng bên trong của các cụm đang được hòa nhập. Nó

có khả năng hơn để khám phá các cụm có hình thù bất kỳ có chất lượng cao hơn CURE.

2.6.3. Các thuật toán phân cụm dựa trên mật độ

Để tìm ra các cụm có mật độ dày, với hình dạng tuỳ ý, các phương pháp phân cụm dựa trên mật độ đã được phát triển, nó kết nối các miền với mật độ đủ cao vào trong các cụm hay phân cụm các đối tượng dựa trên phân bố hàm mật độ.

Chúng ta có các thuật toán phân cụm dựa trên mật độ như : DBSCAN(KDD'96), DENCLUE (KDD'98), CLIQUE (SIGMOD'98)), OPTICS (SIGMOD'99) . . .

1. Thuật toán DBSCAN

Thuật toán DBSCAN (Density – Based Spatial Clustering of Applications with Noise) là một giải thuật phân cụm dựa trên mật độ, được phát triển bởi Ester, Kriegel, Sander và Xu năm 1996. Giải thuật này tăng trưởng các miền với mật độ cao vào trong các cụm và khám phá ra các cụm có hình dạng bất kỳ trong không gian cơ sở dữ liệu có nhiễu.

Ý tưởng cơ bản của phân cụm dựa trên mật độ : Đối với mỗi đối tượng của một cụm, láng giềng trong một bán kính cho trước (ε) (gọi là ε-láng giềng) phải chứa ít nhất một số lượng tối thiểu các đối tượng (MinPts).

Một đối tượng nằm trong một bán kính cho trước (ε) chứa không ít hơn một số lượng tối thiểu các đối tượng láng giềng (MinPts), được gọi là đối tượng nồng cốt (core object) đối với bán kính (ε) và số lượng tối thiểu các điểm (MinPts).

Một đối tượng p là mật độ trực tiếp tiến (directly density-reachable) từ đối tượng q với bán kính ε và số lượng tối thiểu các điểm MinPts trong một

tập các đối tượng D nếu p trong phạm vi ε-láng giềng của q với q chứa ít nhất một số lượng tối thiểu điểm MinPts.

Một đối tượng p là mật độ tiến (density-reachable) từ đối tượng q với bán kính ε và MinPts trong một tập hợp các đối tượng D nếu như có một đối tượng p_1, p_2, …,p_n, p_1=q và p_n=p với $1 \leq i \leq n$, $p_i \in D$ và p_{i+1} là mật độ trực tiếp tiến từ p_i đối với ε và MinPts

Một đối tượng p là mật độ liên kết với đối tượng q đối với ε và MinPts trong một ập đối tượng D nếu như có một đối tượng $o \in D$ để cả p và q là mật độ tiến từ o đối với ε và MinPts.

DBSCAN có thể tìm ra các cụm với hình thù bất kỳ, trong khi đó tại cùng một thời điểm ít bị ảnh hưởng bởi thứ tự của các đối tượng dữ liệu nhập vào. Khi có một đối tượng được chèn vào chỉ tác động đến một láng giềng xác định. Mặt khác, DBSCAN sử dụng tham số ε và MinPts trong thuật toán để kiểm soát mật độ của các cụm. DBSCAN bắt đầu với một điểm tuỳ ý và xây dựng mật độ láng giềng có thể được đối với ε và MinPts. Vì vậy, DBSCAN yêu cầu người dùng xác định bán kính ε của các láng giềng và số các láng giềng tối thiểu MinPts, các tham số này khó mà xác định được tối ưu, thông thường nó được xác định bằng phép chọn ngẫu nhiên hoặc theo kinh nghiệm.

Độ phức tạp của DBSCAN là $O(n^2)$, nhưng nếu áp dụng chỉ số không gian để giúp xác định các láng giềng của một đối tượng dữ liệu thì độ phức của DBSCAN đã được cải tiến là O(nlogn). Thuật toán DBSCAN có thể áp dụng cho các tập dữ liệu không gian lớn đa chiều, khoảng cách Euclide được sử dụng để đo sự tương tự giữa các đối tượng nhưng không hiệu quả đối với dữ liệu đa chiều.

Hinh 2.25 : Hình dạng các cụm được khám phá bởi thuật toán DBSCAN [2]

Thuật toán : DBSCAN khởi tạo điểm p tùy ý và lấy tất cả các điểm liên lạc mật độ từ p tới ε và MinPts. Nếu p là điểm nhân thì thủ tục trên tạo ra một cụm theo ε và MinPts, nếu p là một điểm biên, không có điểm nào liên lạc mật độ từ p và DBSCAN sẽ đi thăm điểm tiếp theo của tập dữ liệu.

Nếu sử dụng giá trị toàn cục ε và MinPts, DBSCAN có thể hoà nhập hai cụm thành một cụm nếu mật độ của hai cụm gần bằng nhau. Giả sử khoảng cách giữa hai tập dữ liệu S1 và S2 được định nghĩa là :

dist(S1, S2) = min{dist(p, q)} {p ∈ S1 và q ∈ S2}.

Thuật toán DBSCAN được mô tả chi tiết như sau:

Modul chương trình chính

```
DBSCAN(SetOfPoints, ε, MinPts)

//SetOfPoints is UNCLASSIFIED

Clusterid:= NextId(NOISE);

FOR i FROM 1 TO SetOfPoints.size DO

    Point := SetOfPoints.get(i);
```

```
        IF PointClId = UNCLASSIFIED THEN
                IF ExpandCluster(SetOfPoints, Point, ClusterId, ε, MinPts) THEN
                        ClusterId.= nextId(ClusterId)
                END IF
        END IF
END FOR
END;
//DBSCAN
```

Thủ tục ExpandCluster

```
ExpandClusster(SetOfPoints, Points, C1Id, ε, MinPts): Boolean;
seeds:= SetOfPoints.regionQuery(Point, ε)
IF seeds.size < MinPts THEN //no core point
        SetOfPoints.changeclId(Point, NOISE),
        RETURN False;
ELSE //all points in seeds are density-reachable from Point
        SetOfPoints.changeClId(seeds, C1Id);
        seeds.delete(Point);
        WHILE seeds <> Empty DO
                currentP:= seeds.first();
```

```
            result:= SetOfPoints.regionQuery(CurrentP, ε);
            IF result.size >= MinPts THEN
                FOR i FROM 1 to result.size DO
                    resultpP:= result.get(i);
                    IF resultp.C1Id IN {UNCLASSIFIED, NOISE}
THEN
                        IF resultp.ClId = UNCLASSIFIED THEN
                            seeds.append(resultP);
                        END IF;
                        SetOfPoints.changeC1Id(resultP, C1Id),
                    END IF; //UNCLASSIFIED or NOISE
                END FOR;
            END IF; //result.size >= Minpts
            seeds.delete(currentP);
        END WHILE; //seeds <> Empty
        RETURN True;
END IF;
END; //ExpandCluster
```

Trong đó SetOfPoints hoặc là tập dữ liệu ban đầu hoặc là cụm được khám phá từ bước trước, C1Id (ClusterId) là nhãn đánh dấu phần tử dữ liệu

nhiễu có thể thay đổi nếu chúng có thể liên lạc mật độ từ một điểm khác trong CSDL, điều này chỉ xảy ra đối với các điểm biên của dữ liệu. Hàm SetOfPoints.get(i) trả về phần tử thứ i của SetOfPoints. Thủ tục SetOfPoints.regionQuery(Point, ε) trả về một danh sách các điểm dữ liệu lân cận với điểm Point trong ngưỡng ε từ tập dữ liệu SetOfPoints. Trừ một số trường hợp ngoại lệ, kết quả của DBSCAN là độc lập với thứ tự duyệt các đối tượng dữ liệu. ε và MinPts là hai tham số toàn cục được xác định bằng thủ công hoặc theo kinh nghiệm. Tham số ε được đưa vào là nhỏ so với kích thước của không gian dữ liệu, thì độ phức tạp tính toán trung bình của mỗi truy vấn là O(logn).

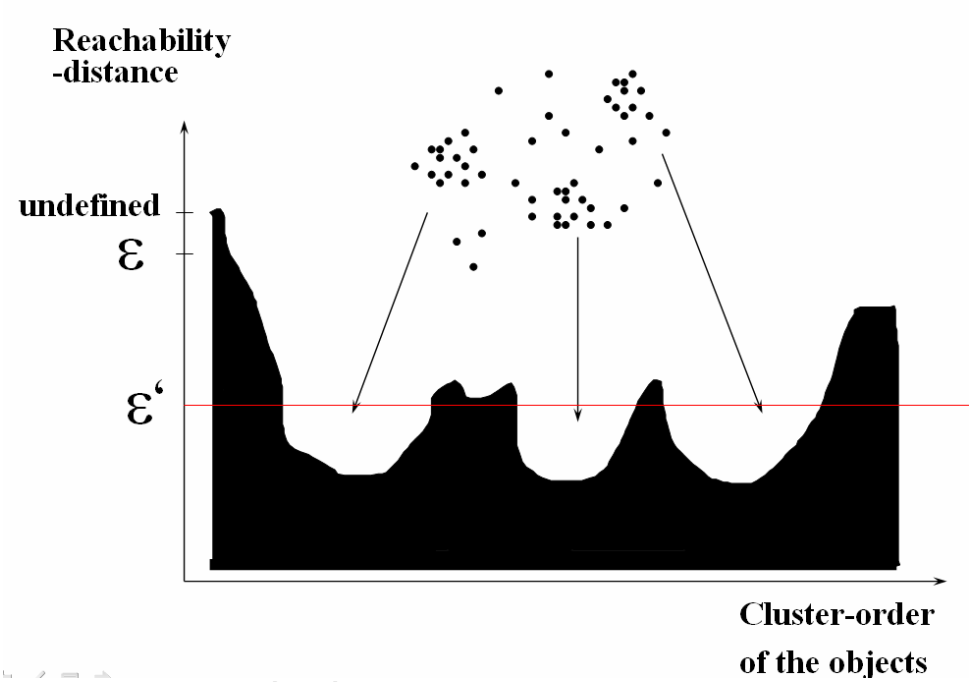
2. Thuật toán OPTICS

Mặc dù giải thuật phân cụm dựa trên mật độ DBSCAN có thể tìm ra cụm các đối tượng với việc lựa chọn các tham số đầu vào như ε và MinPts, người dùng vẫn chịu trách nhiệm lựa chọn các giá trị tham số tốt để tìm ra các cụm chính xác. Trên thực tế, đây là bài toán có sự kết hợp của nhiều giải thuật phân cụm khác. Các thiết lập tham số như vậy tương đối khó, đặc biệt trong thế giới thực, các tập dữ liệu có số chiều cao. Hầu hết các giải thuật rất nhạy với các tham số : các thiết lập có sự khác biệt nhỏ có thể dẫn tới các phân chia dữ liệu rất khác nhau. Hơn nữa, các tập dữ liệu thực số chiều cao thường có phân bố rất lệch, thậm trí ở đó không tồn tại một thiết lập tham số toàn cục cho đầu vào.

Để khắc phục khó khăn này, một phương pháp sắp xếp cụm gọi là OPTICS (Ordering Point To Identify the Clustering Structuer) được phát triển bởi Ankerst, Breunig , Kriegel và Sander năm 1999. nó cải tiến bằng cách giảm bớt các tham số đầu vào. Thuật toán này không phân cụm các điểm dữ liệu mà thực hiện tính toán và sắp xếp trên các điểm dữ liệu theo thứ tự tăng dần nhằm tự động phân cụm dữ liệu và phân tích cụm tương tác hơn là đưa ra phân cụm một tập dữ liệu rõ ràng. Đây là thứ tự mô tả cấu trúc phân dữ liệu

cụm dựa trên mật độ của dữ liệu, nó chứa thông tin tương ứng với phân cụm dựa trên mật độ từ một dãy các tham số được thiết lập và tạo thứ tự của các đối tượng trong cơ sở dữ liệu, đồng thời lưu trữ khoản cách lõi và khoảng cách liên lạc phù hợp của mỗi đối tượng. Hơn nữa, thuật toán được đề xuất rút ra các cụm dựa trên thứ tự thông tin. Như vậy thông tin đủ cho trích ra tất cả các cụm dựa trên mật độ khoảng cách bất kỳ ε' mà nhỏ hơn khoảng cách ε được sử dụng trong sinh thứ tự.

Việc sắp xếp thứ tự được xác định bởi hai thuộc tính riêng của các điểm dữ liệu đó là khoảng cách nhân và khoảng cách liên lạc. Các phép đo này chính là kích thước mà có liên quan đến quá trình của thuật toán DBSCAN, tuy nhiên, chúng được sử dụng để xác định thứ tự của các điểm dữ liệu đã được xắp xếp. Thứ tự dựa trên cơ sở các điểm dữ liệu mà có khoảng cách nhân nhỏ nhất và tăng dần độ lớn. Điều duy nhất về phương pháp này là người sử dụng không phải xác định giá trị ε hoặc MinPts phù hợp.



Hinh 2.26 : Sắp xếp cụm trong OPTICS phụ thuộc vào ε [7]

Thuật toán này có thể phân cụm các đối tượng đã cho với các tham số

đầu vào như ε và MinPts, nhưng nó vẫn cho phép người sử dụng tùy ý lựa chon các giá trị tham số mà sẽ dãn đến khám phá các cụm chấp nhận được. Các thiết lập tham số thường dựa theo kinh nghiệm tập hợp và khó xác định, đặc biệt là với các tập dữ liệu đa chiều.

Tuy nhiên, nó cũng có độ phức tạp thời gian thực hiện như DBSCAN bởi vì có cấu trúc tương đương với DBSCAN : O(nlogn) với n là kích thước của tập dữ liệu. Thứ tự cụm của tập dữ liệu có thể được biểu diễn bằng đồ thị, và được minh họa hình sau, có thể thấy ba cụm, giá trị ε quyết định số cụm.

3. Thuật toán DENCLUDE

DENCLUDE (DENsity -based CLUstEring) do Hinneburg và Keim vào năm 1998 đưa ra cách tiếp cận khác với các thuật toán phân cụm dựa trên mật độ trước đó, cách tiếp cận này xem xét mô hình được sử dụng một công thức toán để mô tả mỗi điểm dữ liệu sẽ ảnh hưởng trong mô hình như thế nào được gọi là hàm ảnh hưởng có thể xem như một hàm mà mô tả ảnh hưởng của điểm dữ liệu với các đối tượng làng giếng của nó. Ví dụ về hàm ảnh hưởng là các hàm parabolic, hàm sóng ngang, hoặc hàm Gaussian.

Như vậy, DENCLUDE là phương pháp dựa trên một tập các hàm phân phố mật độ và được xây dựng ý tưởng chính như sau :

- Ảnh hưởng của mỗi điểm dữ liệu có thể là hình thức được mô hình sử dụng một hàm tính toán, được gọi là hàm ảnh hưởng, mô tả tác động của điểm dữ liệu với các đối tượng láng giềng của nó;
- Mật độ toàn cục của không gian dữ liệu được mô hình phân tích như là tổng các hàm ảnh hưởng của tất cả các điểm dữ liệu;
- Các cụm có thể xác định chính xác bởi việc xác định mật độ cao (density attractors), trong đó mật độ cao là các điểm cực đại hàm mật độ toàn cục.

Sử dụng các ô lưới không chỉ giữ thông tin về các ô lưới mà thực tế nó còn chứa đựng cả các điểm dữ liệu. Nó quản lý các ô trong một cấu trúc truy cập dựa trên cây và như vậy nó nhanh hơn so với một số các thuật toán có ảnh hưởng như DBSCAN. Tuy nhiên, phương pháp này đòi hỏi chọn lựa kỹ lưỡng tham biến mật độ và ngưỡng nhiễu, việc chọn lựa tham số là quan trọng ảnh hưởng tới chất lượng của các kết quả phân cụm.

Định nghĩa : Cho x, y là hai đối tượng trong không gian d chiều ký hiệu là F^d. Hàm ảnh hưởng của đối tượng $y \in F^d$ lên đối tượng x là một hàm $f_B^y : F \to R_0^+$ mà được định nghĩa dưới dạng một hàm ảnh hưởng cơ bản $f_B^y(X) = f_b(x, y)$. Hàm ảnh hưởng có thể là một hàm bất kỳ; cơ bản là xác định khoảng cách của hai vecto d(x, y) trong không gian d chiều, ví dụ như khoảng cách Euclide. Hàm khoảng cách có tính chất phản xạ và đối xứng. Ví dụ về hàm ảnh hưởng như sau [7] :

- Hàm ảnh hưởng sóng ngang : $f_{square}(x,y) = \begin{cases} 0 & \text{if} \quad d(x,y) > \delta \\ 1 & \text{if} \quad d(x,y) \le \delta \end{cases}$

  Trong đó δ là một ngưỡng

- Hàm ảnh hưởng Gaussian : $f_{square}(x,y) = e^{\frac{d(x,y)^2}{2\delta^2}}$

Mặt khác, hàm mật độ tại điểm $x \in F^d$ được định nghĩa là tổng các hàm ảnh hưởng của tất cả các điểm dữ liệu. Cho n là các đối tượng dữ liệu được mô tả bởi một tập véc tơ $D = \{x_1, x_2, \ldots, x_n\} \in F^d$ hàm mật độ được định nghĩa như sau :

$$F_B^D(x) = \sum_{i=1}^{n} F_B^{x(i)}(x)$$

Hàm mật độ được thành lập dựa trên ảnh hưởng Gauss được xác định như sau :

$$F_{Gauss}^D(d) = \sum_{i=1}^{n} e^{\frac{d(x,x_i)^2}{2\delta^2}}$$

DENCLUE phụ thuộc nhiều vào ngưỡng nhiễu và tham số mật độ

nhưng DENCLUE có các lợi thế chính được so sánh với các thuật toán phân cụm khác sau đây :

- Có cơ sở toán học vững chắc và tổng quát hóa các phương pháp phân cụm khác, bao gồm các phương pháp phân cấp, dựa trên phân hoạch.
- Có các đặc tính phân cụm tốt cho các tập dữ liệu với số lượng lớn và nhiễu.
- Cho phép các cụm có hình dạng bất kỳ trong tập dữ liệu đa chiều được mô tả trong công thức toán.

Độ phức tạp tính toán của DENCLUDE là O(nlogn). Các thuật toán dựa trên mật độ không thực hiện kỹ thuật phân mẫu trên tập dữ liệu như trong các thuật toán phân cụm phân hoạch, vì điều này có thể làm tăng thêm độ phức tạp đã có sự khác nhau giữa mật độ của các đối tượng trong mẫu với mật độ của toàn bộ dữ liệu.

2.6.4. Các thuật toán phân cụm dựa vào lưới

Một tiếp cận dựa trên lưới dùng cấu trúc dữ liệu lưới đa phân giải. Trước tiên nó lượng tử hóa không gian vào trong một số hữu hạn các ô mà đã hình thành nên cấu trúc lưới, sau đó thực hiện tất cả các thao tác trong cấu trúc lưới đó. Thuận lợi chính của tiếp cận này là thời gian xử lý nhanh, điển hình là độc lập của số lượng các đối tượng dữ liệu nhưng độc lập chỉ trên số lượng các ô trong mỗi chiều trong không gian lượng tử hóa.

Phân cụm dữ liệu dựa trên lưới bao gồm STING khảo sát thông tin thống kê được lưu trữ trong các ô lưới; WaveCluster các cụm đối tượng sử dụng phương pháp biến đổi wavelet; CLIQUE miêu tả một tiếp cận dựa trên lưới và mật độ cho phân cụm trong không gian dữ liệu số chiều cao.

1. Thuật toán STING

STING (STatistical INformation Grid) do Wang, Yang và Munz phát triển năm 1997, là kỹ thuật phân cụm đa phân giải dựa trên lưới, trong đó vùng không gian dữ liệu được phân rã thành số hữu hạn các ô chữ nhât, điều này có ý nghĩa là các ô lưới được hình thành từ các ô lưới con để thực hiện phân cụm. Có nhiều mức của các ô chữ nhật tương ứng với các mức khác nhau của phân giải trong cấu trúc lưới, và các ô này hình thành cấu trúc phân cấp : mỗi ô ở mức cao được phân hoạch thành các ô nhỏ ở mức thấp hơn tiếp theo trong cấu trúc phân cấp. Các điểm dữ liệu được nạp từ CSDL, giá trị của các tham số thống kê cho các thuộc tính của đối tượng dữ liệu trong mỗi ô lưới được tính toán từ dữ liệu và lưu trữ thông qua các tham số thống kê ở các ô mức thấp hơn (điều này giống với cây CF). Các giá trị của các tham số thống kê gồm : số trung bình – mean, số tối đa – max, số tối thiểu – min, số đếm –count , độ lệch chuẩn –s,…

Các đối tượng dữ liệu lần lượt được chèn vào lưới và các tham số thống kê ở trên được tính trực tiếp thông qua các đối tượng dữ liệu này. Các truy vấn không gian được thực hiện bằng cách xét các ô thích hợp tại mỗi mức phân cấp. Một truy vấn không gian được xác định như là một thông tin khôi phục lại của dữ liệu không gian và các quan hệ của chúng. STING có khả năng mở rộng cao, nhưng do sử dụng phương pháp đa phân giải nên nó phụ thuộc chặt chẽ vào trọng tâm của mức thấp nhất. Đa phân giải là khả năng phân rã tập dữ liệu thành các mức chi tiết khác nhau. Khi hòa nhập các ô của cấu trúc lưới để hình thành các cụm, nó không xem xét quan hệ không gian giữa các nút của mức con không được hòa nhập phù hợp (do chúng chỉ tương ứng với các cha của nó) và hình dạng của các cụm dữ liệu khám phá là isothetic, tất cả ranh giới của các cụm có các biên ngang và dọc, theo biên của các ô và không có đường biên chéo được phát hiện ra.

Đánh giá thuật toán STING

- Ưu điểm:

- Tính toán dựa trên lưới là truy vấn độc lập vì thông tin thống kê được bảo quản trong mỗi ô đại diện nên chỉ cần thông tin tóm tắt của dữ liệu trong ô chứ không phải là dữ liệu thực tế và không phụ thuộc vào câu truy vấn.
- Cấu trúc dữ liệu lưới thuận tiện cho quá trình xử lý song song và cập nhật liên tục.
- Duyệt toàn bộ CSDL một lần để tính toán các đại lượng thống kê cho mỗi ô, nên nó hiệu quả và do đó độ phức tạp thời gian để tạo các cụm xấp xỉ O(n), trong đó n là tổng số các đối tượng. Sau khi xây dựng cấu trúc phân cấp, thời gian xử lý cho các truy vấn là O(g), trong đó g là tổng số ô lưới ở mức thấp (g<<n).

  - Nhược điểm:

- Trong khi sử dụng cách tiếp cận đa phân giải để thực hiện phân tích cụm chất lượng của phân cụm STING hoàn toàn phụ thuộc vào tính chất hộp ở mức thấp nhất của cấu trúc lưới.
- Nếu tính chất hộp là mịn, dẫn đến chi phí thời gian xử lý tăng, tính toán trở nên phức tạp và nếu mức dưới cùng là quá thô thì nó có thể làm giảm bớt chất lượng và độ chính xác của phân tích cụm.

Thuật toán STING :

Bước 1. Xác định tầng để bắt đầu .

Bước 2. Với mỗi cái của tầng này, tính toán khoảng tin cậy (hoặc ước lượng khoảng) của xác suất mà ô này liên quan tới truy vấn.

Bước 3. Từ khoảng tin cậy của tính toán trên, gán nhãn cho là có liên quan hoặc không liên quan.

Bước 4. Nếu lớp này là lớp cuối cùng, chuyển sang Bước 6; nếu

khác thì chuyển sang Bước 5.

Bước 5. Duyệt xuống dưới của cấu trúc cây phân cấp một mức. Chuyển sang Bước 2 cho các ô mà hình thành các ô liên quan của lớp có mức cao hơn.

Bước 6. Nếu đặc tả được câu truy vấn, chuyển sang bước 8; nếu không thì chuyển sang bước 7.

Bước 7. Truy lục lại dữ liệu vào trong các ô liên quan và thực hiện xử lý. Trả lại kết quả phù hợp yêu cầu của truy vấn. Chuyển sang Bước 9.

Bước 8. Tìm thấy các miền có các ô liên quan. Trả lại miền mà phù hợp với yêu cầu của truy vấn. Chuyển sang bước 9.

Bước 9. Dừng

2. Thuật toán WaveCluster

Thuật toán WaveCluster do Sheikholeslami, Chatterjee và Zhang đề xuất năm 1998, là phương pháp gần giống với STING, tuy nhiên thuật toán sử dụng phép biến đổi dạng sóng để tìm ô đặc trong không gian. Đầu tiên kỹ thuật này tóm tắt dữ liệu bằng việc tận dụng cấu trúc dạng lưới đa chiều lên trên không gian dữ liệu. Tiếp theo nó sử dụng phép biến đổi dạng sóng để biến đổi không gian có đặc trưng gốc, tìm kiếm ô đặc trong không gian đã được biến đổi. Phương pháp này là phức tạp với các phương pháp khác chính là ở phép biến đổi.

Ở đây, mỗi ô lưới tóm tắt thông tin các điểm của một nhóm ánh xạ vào trong ô. Đây là thông tin tiêu biểu thích hợp đưa vào bộ nhớ chính để sử dụng phép biến đổi dạng sóng đa phân giải và tiếp theo là phân tích cụm. Một phép biến đổi dạng sóng là kỹ thuật dựa trên cơ sở xử lý tín hiệu và xử lý ảnh bằng phân tích tín hiệu với tần số xuất hiện trong bộ nhớ chính. Bằng việc thực

hiện một loạt các phép biến đổi ngược phức tạp cho nhóm này, nó cho phép các cụm trong dữ liệu trở thành rõ ràng hơn. Các cụm này có thể được xác định bằng tìm kiếm ô đặc trong vùng mới.

Phương pháp này phức tạp, nhưng lại có những lợi thế :

Cung cấp cụm không giám sát, khử nhiễu các thông tin bên ngoài biên của cụm. Theo cách đó, vùng đặc trong không gian đặc trưng gốc hút các điểm ở gần và ngăn chặn các điểm ở xa. Vì vậy, các cụm tự động nổi bật và làm sạch khu vực xung quanh nó, do đó các kết quả tự động loại phần tử ngoại lai.

- Đa phân giải là thuộc tính hỗ trợ dò tìm các cụm có các mức biến đổi chính xác.
- Thực hiện nhanh với độ phức tạp của thuật toán là O(n), trong đó n là số đối tượng trong CSDL. Thuật toán có thể thích hợp với xử lý song song.
- Xử lý tập dữ liệu lớn có hiệu quả, khám phá các cụm có hình dạng bất kỳ, xử lý phần tử ngoại lai, mẫn cảm với thứ tự vào, và không phụ thuộc vào các tham số vào như số các cụm hoặc bán kính láng giềng.

3. Thuật toán CLIQUE

Trong không gian đa chiều, các cụm có thể tồn tại trong tập con của các chiều hay còn gọi là không gian con. Thuật toán CLIQUE là thuật toán hữu ích cho phân cụm dữ liệu không gian đa chiều trong các CSDL lớn thành các không gian con. Thuật toán này bao gồm các bước :

- Cho n là tập lớn của các điểm dữ liệu đa chiều; không gian dữ liệu thường là không giống nhau bởi các điểm dữ liệu. Phương pháp này xác định những vùng gần, thưa và "đặc" trong không gian dữ liệu nhất định, bằng cách đó phát hiện ra toàn thể phân bố mẫu của tập dữ liệu.

- Một đơn vị là dày đặc nếu phần nhỏ của tất cả các điểm dữ liệu chứa trong nó vượt quá tham số mẫu đưa vào. Trong thuật toán CLIQUE, cụm được định nghĩa là tập tối đa liên thông các đơn vị dày đặc.

Các đặc trưng của CLIQUE

- Tự động tìm kiếm không gian con của không gian đa chiều, sao cho mật độ đặc của các cụm tồn tại trong không gian con.
- Mẫn cảm với thứ tự của dữ liệu vào và không phù hợp với bất kỳ quy tắc phân bố dữ liệu nào.
- Phương pháp này tỷ lệ tuyến tính với kích thước vào và có tính biến đổi tốt khi số chiều của dữ liệu tăng.

Nó phân hoạch tập dữ liệu thành các hình hộp chữ nhật và tìm các hình hộp chữ nhật đặc, nghĩa là các hình hộp này chứa một số các đối tượng dữ liệu trong số các đối tượng láng giềng cho trước. Hợp các hình hộp này tạo thành các cụm dữ liệu. Tuy nhiên, CLINQUE được bắt đầu bằng cách tiếp cận đơn giản do đó chính xác của kết quả phân cụm có thể bị ảnh hưởng dẫn tới chất lượng của các phương pháp này có thể giảm.

Phương pháp bắt đầu nhận dạng các ô đặc đơn chiều trong không gian dữ liệu và tìm kiếm phân bố của dữ liệu, tiếp đến CLINQUE lần lượt tìm các hình chữ nhật 2 chiều, 3 chiều,…., cho đến khi hình hộp chữ nhật đặc k chiều được tìm thấy, độ phức tạp tính toán của CLIQUE là O(n)

2.6.5. Các thuật toán phân cụm dựa trên mô hình

Điển hình trong phương pháp tiếp cận theo phân cụm dựa trên mô hình là các thuật toán như : EM, COBWEB, CLASSIT, AutoClass (Cheeseman and Stutz, 1996) . . .

1. Thuật toán EM

Thuật toán EM được xem như là thuật toán dựa trên mẫu hoặc là mở rộng của thuật toán K-means. Thật vậy, EM gán các đối tượng cho các cụm đã cho theo xác suất phân phỗi thành phần của đối tượng đó. Phân phối xác suất thường được sử dụng là phân phối xác suất Gaussian với mục đích là khám phá lặp các giá trị tốt cho các tham số của nó bằng hàm tiêu chuẩn là hàm logarit khả năng của đối tượng dữ liệu, đây là hàm tốt để mô hình xác suất cho các đối tượng dữ liệu. EM có thể khám phá ra nhiều hình dạng cụm khác nhau, tuy nhiên do thời gian lặp của thuật toán khá nhiều nhằm xác định các tham số tốt nên chi phí tính toán của thuật toán khá cao. Đã có một số cải tiến được đề xuất cho EM dựa trên các tính chất của dữ liệu : có thể nén, có thể sao lưu trong bộ nhớ và có thể hủy bỏ. Trong các cải tiến này, các đối tượng bị hủy bỏ khi biết chắc chắn được nhãn phân cụm của nó, chúng được nén khi không loại bỏ và thuộc về một cụm quá lớn so với bộ nhớ và chúng sẽ được lưu lại trong các trường hợp còn lại.

Thuật toán được chia thành hai bước và quá trình đó được lặp lại cho đến khi vấn đề được giải quyết :

$$-E : \mu \to a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu} h, \qquad b = \frac{\mu}{\frac{1}{2}+\mu} h$$

$$-M : a, b \to \mu = \frac{a+b}{6(b+c+d)}$$

1. Khởi tạo tham số :

$$\lambda_0 = \left\{ \mu_1^{(0)}, \mu_2^{(0)}, ..., \mu_K^{(0)}, p_1^{(0)}, p_2^{(0)}, ..., p_k^{(0)} \right\}$$

2. Bước E :

$$P(\omega_j \left| x_k, \lambda_t \right.) = \frac{P(x_k \left| \omega_j, \lambda_t \right.) P(\omega_j, \lambda_t)}{P(x_k, \lambda_t)} = \frac{P(x_k \left| \omega_i, \lambda_i^{(t)}, \sigma^2 \right.) P_i^{(t)}}{\sum_k P(x_k \left| \omega_i, \lambda_i^{(t)}, \sigma^2 \right.) P_j^{(t)}}$$

3. Bước M :

$$\mu_i^{t+1} = \frac{\sum_k P(\omega_i \left| x_k, \lambda_t \right) x_k}{\sum_k P(\omega_i \left| x_k, \lambda_t \right)}$$

$$p_i^{(t+1)} = \frac{\sum_k P(\omega_i \left| x_k, \lambda_t \right)}{R}$$

4. Lặp lại bước 2, 3 cho đến khi đạt kết quả

2. Thuật toán COBWEB

COBWEB là cách tiếp cận để biểu diễn các đối tượng dữ liệu theo kiểu cặp thuộc tính – giá trị. COBWEB thực hiện bằng cách tạo cây phân lớp, tương tự như khái niệm của BIRCH, tuy nhiên cấu trúc cây khác nhau. Mỗi nút của cây phân lớp là đại diện cho khái niệm của đối tượng dữ liệu và tất cả các điểm mà ở dưới lớp đó là cùng thuộc một nút. COBWEB sử dụng công cụ phân loại để quản lý cấu trúc cây. Từ đó các cụm hình thành dựa trên phép đo độ tương tự mà phân loại giữa tương tự và phi tương tự, cả hai có thể mô tả phân chia giá trị thuộc tính giữa các nút trong lớp. Cấu trúc cây cũng có thể mô tả phân chia giá trị thuộc tính giữa các nút trong lớp. Cấu trúc cây cũng có thể được hợp nhất hoặc phân tách khi chèn một nút mới vào cây. Có hai phương pháp cải tiến cho COBWEB là CLASSIT và AutoClass.

2.7. Kết luận

Chương này đề cập đến một số phương pháp phân cụm truyền thống và một số cách cải tiến phân cụm truyền thống, các ưu nhược điểm của từng phương pháp đối với từng loại dữ liệu. Qua đó ta có thể thấy được khả năng phân cụm của từng phương pháp, khả năng áp dụng vào các bài toán thực tiễn.

Chương 3. Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh

3.1. Đặt vấn đề

Giả sử bạn nằm trong Ban giám hiệu một trường Trung học Phổ thông nào đó. Sau khi có điểm cuối học kì, cuối năm học … bạn nhìn bảng điểm của lớp, của trường … và muốn đưa ra một vài kết luận nào đó về tình hình học tập của học sinh trường mình phụ trách. Bạn sẽ làm thế nào?

Thứ nhất, tính điểm trung bình. Đây là một trong những cách đơn giản nhất. Dựa vào điểm trung bình có thể đưa ra vài nhận xét về tình hình học tập. Giả sử năm nay điểm trung bình môn Toán của khối lớp 12 là 7.5, của khối lớp 11 là 7.6, như vậy có thể nhận xét "nóng vội" rằng năm nay học sinh khối lớp 11 học môn Toán tốt hơn học sinh khối lớp 12 và cả 2 khối lớp năm nay đều học khá môn Toán …

Tất nhiên những nhận xét này là "nóng vội" vì có thể có những học sinh điểm rất cao, ngược lại cũng có điểm rất thấp. Rõ ràng tính trung bình rất đơn giản và "mạnh mẽ" (từ điểm của vài trăm học sinh một khối, tức là vài trăm mẫu dữ liệu khác nhau, chuyển thành một điểm trung bình duy nhất, và có thể dựa vào điểm trung bình để nhận xét về điểm của vài trăm học sinh), nhưng vì thế mà nó làm mất thông tin về phân bố của dữ liệu.

Thứ hai, tính kì vọng và phương sai. Một người học lớp thống kê cơ bản sẽ biết cách tính kì vọng và phương sai của dữ liệu. Trong trường hợp này thì kì vọng chính là điểm trung bình cộng. Dựa vào kì vọng và phương sai, ta sẽ có thể đưa ra những nhận xét sâu sắc và ít "nóng vội" hơn. Chẳng hạn nếu khối lớp 11 có điểm kì vọng môn Toán là 7.6 và phương sai 2.5, khối 12 có điểm kì vọng 7.0 nhưng phương sai 1.5, thì có thể kết luận là nhìn chung khối lớp 11 học Toán tốt hơn khối lớp 12, nhưng khói lớp 12 học "đều" môn Toán

hơn khối lớp 11 (nghĩa là không có chênh lệch quá lớn giữa các học sinh)… và có thể có hàng chục kết luận khác nữa.

Thứ ba, đếm số lượng điểm trong ngưỡng nào đó. Chẳng hạn ta có thể nói: trong 500 học sinh thì có 400 học sinh đạt điểm trên 5.0 và 100 học sinh dưới 5.0. Như vậy có thể hiểu khối lớp có 4/5 học sinh trên trung bình. Một cách chi tiết hơn, ta có thể nói: trong 400 học sinh trên trung bình thì có 100 học sinh là trên 8.0, 200 học sinh là từ 6.5 đến cận 8.0, và 100 học sinh là từ 5.0 đến cận 6.0. Như vậy có thể hiểu sâu sắc hơn rằng khối lớp có 1/5 học sinh giỏi, 2/5 học sinh khá v.v….

Vài dòng nêu trên để chúng ta thấy được rằng việc phân tích, đánh giá kết quả học tập của học sinh trong nhà trường không phải là chuyện đơn giản. Nó đòi hỏi Ban giám hiệu nhà trường, các nhà quản lý giáo dục có một sự đầu tư, nghiên cứu, tìm tòi và sáng tạo … nhằm đưa ra được các phân tích, đánh giá đúng đắn nhất, chính xác nhất về kết quả học tập của học sinh từ đó đề ra định hướng, hoạch định cho nhà trường trong việc: đầu tư bồi dưỡng giáo viên bộ môn còn yếu, phát hiện học sinh giỏi để bồi dưỡng, học sinh kém để phụ đạo, có kế hoạch tăng giờ, tăng tiết, định hướng nghề nghiệp cho học sinh dựa trên sở thích, năng khiếu môn học v.v…

Bản thân người viết đề tài cũng đã từng trải qua những khó khăn, vất vả trên và hiện nay ở cấp độ Sở Giáo dục và Đào tạo càng thêm trăn trở nên cố gắng vận dụng những kiến thức được học tại lớp Cao học ngành Công nghệ Thông tin của Trường Đại học Lạc Hồng để đưa ra một công cụ hỗ trợ phân tích điểm số của học sinh ra đa dạng hơn, đa chiều hơn, nhiều góc độ hơn … nhằm giúp cho Ban giám hiệu, các nhà quản lý giáo dục có thêm cơ sở để đánh giá đúng đắn nhất, chính xác nhất .. . về tình hình học tập của học sinh, hoạt động giảng dạy của giáo viên đó là chương trình "Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh".

3.2. Cơ sở lý luận, khoa học và thực tiễn

3.2.1. Cơ sở lý luận

- Nhìn chung cách đánh giá, xếp loại của Bộ Giáo dục và Đào tạo hiện nay phản ánh được mục đích, yêu cầu … việc phân loại học sinh, khuyến khich học sinh học đều các môn góp phần giáo dục học sinh phát triển toàn diện.

- Bộ Giáo dục và Đào tạo quy định "Đánh giá chất lượng giáo dục toàn diện đối với học sinh sau mỗi học kỳ, mỗi năm học nhằm thúc đẩy học sinh rèn luyện, học tập để không ngừng tiến bộ" (Điều 2 của Qui chế đánh giá, xếp loại học sinh THCS & THPT ban hành theo QĐ 40/2006-BGD-ĐT ngày 05/10/2006 của Bộ Giáo dục và Đào tạo).

- "Đánh giá kết quả học tập của học sinh theo mục tiêu chương trình giáo dục phổ thông nhằm góp phần điều chỉnh việc thực hiện chương trình giáo dục phổ thông hiện hành và tạo cơ sở thực tiễn cho việc phát triển chương trình giáo dục phổ thông tiếp theo" "Xác định những nhân tố tác động đến kết quả học tập nhằm cung cấp thông tin góp phần điều chỉnh các chính sách giáo dục hiện hành và xây dựng những chính sách mới để phát triển sự nghiệp giáo dục phổ thông. (Điều 3 của Thông tư Quy định về đánh giá định kỳ quốc gia kết quả học tập của học sinh phổ thông)

Như vậy, việc đánh giá xếp loại học tập của học sinh không phải chỉ nhằm "thúc đẩy học sinh rèn luyện, học tập để không ngừng tiến bộ" mà còn làm thông tin, tiền đề cho việc "điều chỉnh việc thực hiện chương trình giáo dục phổ thông hiện hành và tạo cơ sở thực tiễn cho việc phát triển chương trình giáo dục phổ thông tiếp theo". Đây là vấn đề mà nhà trường, ngành giáo dục và xã hội đang rất quan tâm.

3.2.2. Cơ sở thực tiễn

- Đối với học sinh học sinh ở trường chuyên, trường thi tuyển đầu vào (phần lớn xếp loại học lực là Khá và Giỏi) hay học sinh trường ngoài công lập, trường vừa học vừa làm (phần lớn xếp loại học lực là Trung bình và Yếu) thì cách đánh giá xếp loại học lực theo 5 loại : Giỏi, Khá, Trung bình, Yếu, Kém … chưa phản ánh hết tình hình học tập của học sinh.
- Chủ trương đánh giá xếp loại của Bộ Giáo dục và Đào tạo là đánh giá toàn diện các môn từ đó dễ bỏ sót học sinh có năng khiếu hoặc học sinh giỏi nhưng lệch môn. (vì nếu có một môn Văn hoặc Toán có ĐTB dưới 5 thì không thể xếp loại Giỏi, Khá nhưng những học sinh này vẫn có thể đậu vào Đại học nếu được định hướng đúng).
- Phần lớn các trường THPT trong tỉnh Đồng Nai đều đã sử dụng Chương trình Quản lý điểm học sinh nên có một cơ sở dữ liệu điểm qua các năm tương đối đầy đủ.

3.2.3. Cơ sở khoa học

Trong các thuật toán phân cụm đã tìm hiểu ở Chương 2 thì thuật toán k-means có tốc độ tương đối nhanh, thích hợp với dữ liệu số khi không có phần tử ngoại lai hay giá trị nhiễu. Và thật vậy, dữ liệu điểm của học sinh trong trường phổ thông hiện nay đáp ứng tốt yêu cầu của thuật toán k-means (dữ liệu số, điểm số trãi dài chỉ từ 0-10 và không có phần tử nhiễu…). Từ đó, người viết mạnh dạn đề xuất sử dụng thuật toán k-means để giải quyết bài toán đưa ra.

3.3. Chương trình ứng dụng

3.3.1. Mục đích chương trình

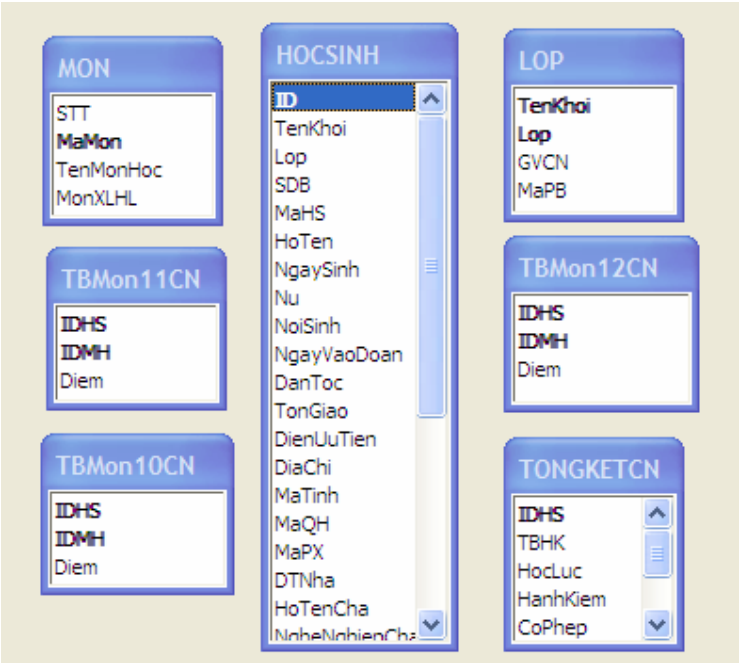
Dựa trên điểm trung bình của từng môn học, điểm trung bình từng học kỳ, cả năm của từng học sinh, từng lớp, từng khối gom cụm dữ liệu nhằm

phân tích điểm số để đưa ra cái nhìn đa dạng hơn, đa chiều hơn, nhiều góc độ khác nhau hơn về điểm số giúp cho Ban giám hiệu, các nhà quản lý giáo dục có thêm cơ sở để đánh giá đúng đắn nhất, chính xác nhất ... về tình hình học tập của học sinh, hoạt động giảng dạy của giáo viên từ đó đề ra định hướng, hoạch định cho nhà trường trong việc nâng cao chất lượng giáo dục.

3.3.2. Cơ sở dữ liệu

Trong đó : bao gồm các table

- HocSinh (ID, TenKhoi, Lop, SBD, MaHS . ..) lưu trữ Lý lịch học sinh
- Lop(TenKhoi, Lop, GVCN, MaPB) lưu trữ Danh sách các lớp.
- Mon(STT,MaMon, TenMonHoc..) lưu trữ các môn học trong trường.
- TBMon10CN(IDHS, IDMH, Diem) lưu trữ điểm môn học khối lớp 10
- TBMon11CN(IDHS, IDMH, Diem) lưu trữ điểm môn học khối lớp 11



Hinh 3.1 : Các table sử dụng trong chương trình

- TBMon12CN(IDHS, IDMH, Diem) lưu trữ điểm môn học khối lớp 12
- TONGKETCN(IDHS, TBHK, HocLuc, Hanhkiem ...) lưu trữ điểm trung bình, xếp loại của học sinh toàn trường

3.3.3. Cài đặt chương trình và sử dụng

1. Cài đặt chương trình

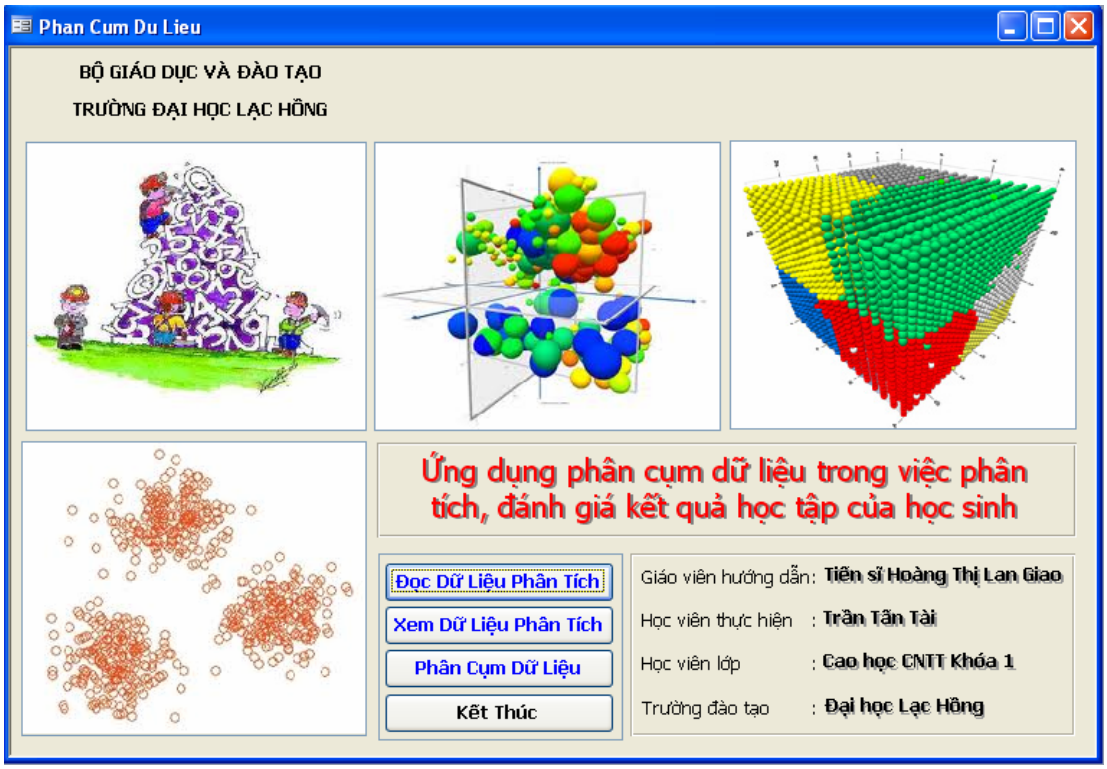
- Tải hoặc chép tập tin CTPhanCum.mdb về máy
- Chép vào thư mục chứa tập tin data.mdb của nhà trường đã có.

2. Khởi động chương trình

- Double click vào tập tin CTPhanCum.mdb để chạy

3.4. Các chức năng chính của chương trình

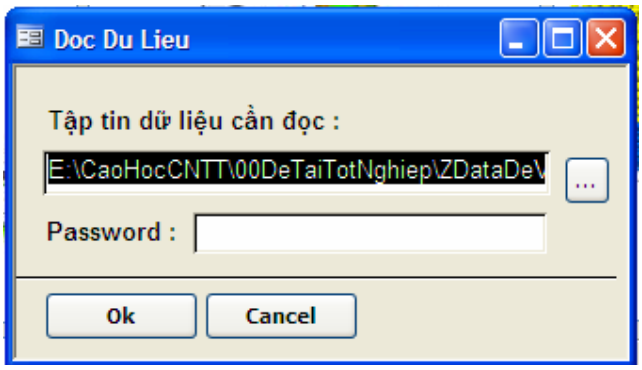
3.4.1. Màn hình khởi động



Hinh 3.2 : Màn hình chính của chương trình

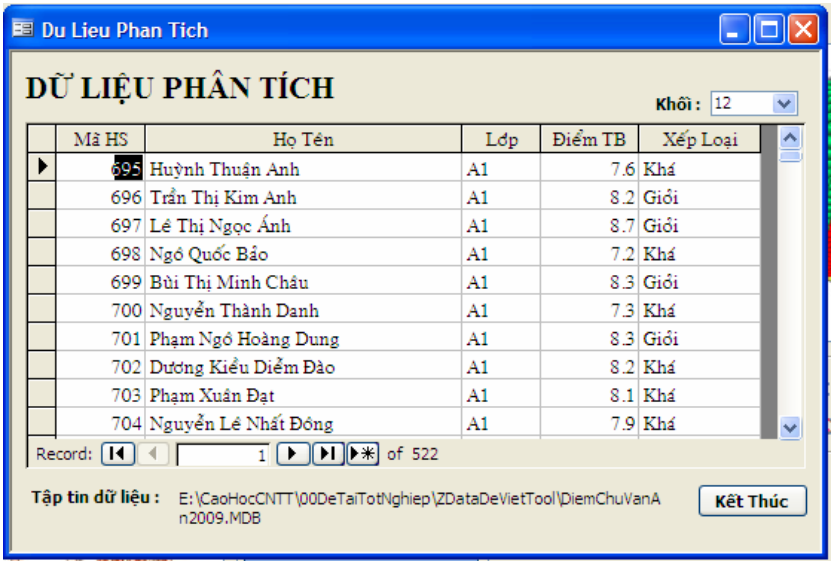
3.4.2. Đọc dữ liệu phân tích : liên kết với tập tin cần phân tích

Tập tin dữ liệu cần đọc : chọn tên tập tin dữ liệu cần liên kết



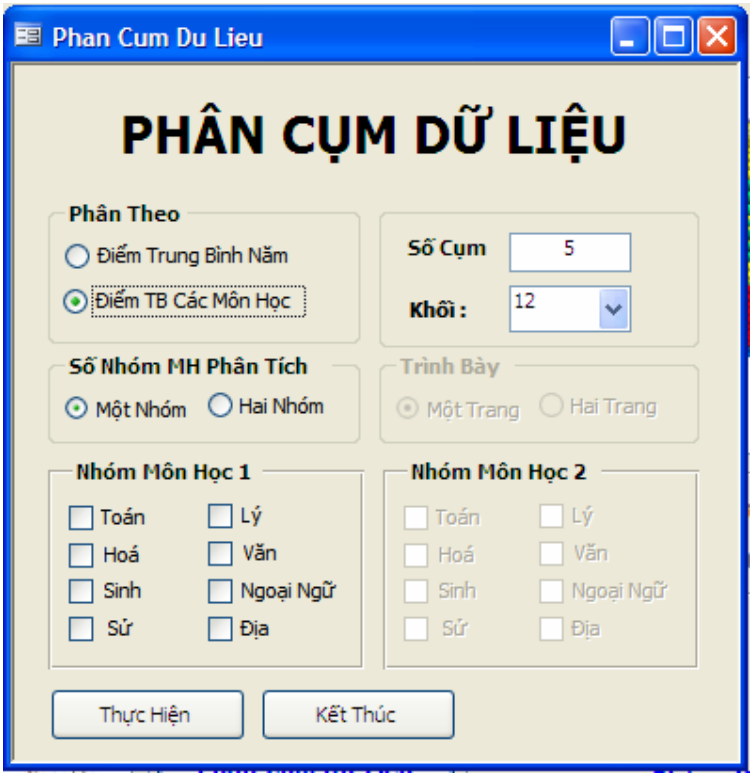
Hinh 3.3 : Màn hình chọn tập tin dữ liệu cần phân tích

3.4.3. Xem dữ liệu phân tích : xem nội dung tập tin cần phân tích



Hinh 3.4 : Màn hình xem trước dữ liệu sẽ được phân tích

3.4.4. Phân cụm dữ liệu : thực hiện việc phân cụm dữ liệu



Hình 3.5 : Màn hình các mục chọn phân cụm

Gồm các chức năng chính :

- Phân cụm theo điểm trung bình năm
- Phân cụm theo điểm trung bình các môn học

Trong đó cho phép chọn lại : Khối lớp cần phân tích hoặc số cụm cần phân tích

- Trong phần Phân cụm theo điểm trung bình các môn học cho phép chọn một môn hoặc nhiều môn.
- Hai môn hoặc hai nhóm môn cùng lúc.

3.4.5. Một số đoạn code chính trong chương trình :

```
'Đọc số cụm cần phân chia .
bSocum = Me.socum.Value
……
' Đọc dữ liệu từ file dữ liệu  :
SQL = "SELECT QTBMON12CN.*, HOCSINH.TenKhoi, HOCSINH.lop, HOCSINH.hoten " & _
    "FROM HOCSINH INNER JOIN QTBMON12CN ON HOCSINH.ID = QTBMON12CN.IDHS "& _
    "WHERE (((HOCSINH.TenKhoi)=" & bkhoi & "));"
    Tentable = "QTBMON12CN"
Set dbf = CurrentDb
Set db = dbf.OpenRecordset(SQL, dbOpenDynaset)
i = 0
Do While Not db.EOF
   i = i + 1
   AHS(i, 1) = db!idhs
   AHS(i, 2) = db!lop
   AHS(i, 3) = db!hoten
   For j = 1 To SoMonHoc_
      AHS(i, 3 + j) = db.Fields(AMonhoc(j)).Value
   Next
   db.MoveNext
Loop
db.Close
```

AHS : là array 2 chiều có số dòng bằng số mẫu tin và số cột là số môn chọn cần phân tích + 3

```
' Khởi tạo trọng tâm các cụm ban đầu
SqlB = "SELECT DISTINCT TOP " & bSocum & " "
For j = 1 To SoMonHoc_
   SqlB = SqlB & AMonhoc(j) & ","
Next
SqlB = Left(SqlB, Len(SqlB) - 1)
SqlB = SqlB & " FROM " & Tentable & ";"
Set db = dbf.OpenRecordset(SqlB, dbOpenDynaset)
i = 0
```

Trọng tâm cụm ban đầu được chọn bằng cách lấy ra **k** mẫu tin đầu tiên có giá trị không trùng lắp.

```
Do While Not db.EOF
   i = i + 1
   For j = 1 To SoMonHoc_
      ATAM(i, j) = db.Fields(j - 1).Value
   Next
   db.MoveNext
Loop
db.Close
```

ATAM : array 2 chiều lưu trữ trọng tâm các cụm, có số dòng bằng số cụm, số cột bằng số môn học chọn phân tích.

```
' Phân bổ các phần tử vào các cụm đã có.
kthuc = False
solanlap = 0
Do While Not kthuc
   solanlap = solanlap + 1
   ' Lưu lại kết quả phân cụm đang có
If solanlap > 1 then
   For i = 1 To SoMT
      For j = 1 To bSocum
         GB(i, j) = GA(i, j)
      Next
   Next
End if
   ' Tính lại trọng tâm các cụm
If solanlap > 1 then
   For j = 1 To bSocum
      For m = 1 To SoMonHoc_
         k = 0
         tong = 0
         For i = 1 To SoMT
            If GA(i, j) <> "0" Then
               k = k + 1
               tong = tong + AHS(i, 3 + m)
            End If
         Next
         If k <> 0 Then
            ATAM(j, m) = tong / k
         End If
```

GA,GB : array 2 chiều lưu trữ kết quả phân cụm, có số dòng bằng số học sinh, số cột bằng số cụm cần phân chia.

Trọng tâm các cụm được tính bằng bình quân giá trị các phần tử trong cụm

```
        Next
    Next
End if
    '----------------------
```

```
    kthuc = True
    'Phân bổ các phần tử vào các cụm
    For i = 1 To SoMT
        For j = 1 To bSocum
            tong = 0
            For m = 1 To SoMonHoc_
                tong = tong + ((AHS(i, 3 + m) - ATAM(j, m)) ^ 2)
            Next
            D(j) = Sqr(tong)
        Next
        min = 1
        For j = 1 To bSocum
            GA(i, j) = "0"
            If D(j) < D(min) Then min = j
        Next
        GA(i, min) = AHS(i, 1)
        ' Gán chỉ số cụm vào phần tử I của AHS
        AHS(i, 3 + SoMonHoc_ + 1) = min
    Next

    ' So sánh kết quả phân cụm
For i = 1 To SoMT
        For j = 1 To bSocum
            If GA(i, j) <> GB(i, j) Then kthuc = False
        Next
    Next
Loop
SoHS_ = SoMT
SolanLap_ = solanlap
```

Tính khoảng cách Euclide của phần tử **i** đến cụm **j**.

Tìm chỉ số cụm có khoảng cách của phần tử **i** đến cum đó là nhỏ nhất và gán phần tử **i** vào cụm đó.

So sánh kết quả phân cụm hiện tại (GA) với kết quả đã lưu trước đó (GB) .

```
' Ghi bảng phân cụm tổng hợp vào tập tin dữ liệu
Set dbf = CurrentDb
dbf.Execute "Delete * from kqcum_mon;"
Set db = dbf.OpenRecordset("kqcum_mon", dbOpenDynaset)
For j = 1 To bSocum
   SoPT = 0
   For i = 1 To SoMT
      If GA(i, j) <> "0" Then SoPT = SoPT + 1
   Next
   db.AddNew
   db!TTCum = j
   db!SoPT = SoPT
   tong = 0
   For m = 1 To SoMonHoc_
      db.Fields(1 + m).Value = Round(ATAM(j, m), 2)
      tong = tong + ATAM(j, m)
   Next
   db!tyle = Round((SoPT / SoMT) * 100, 2)
   db!tongdiem = Round(tong, 2)
   db.Update
Next
db.Close
```

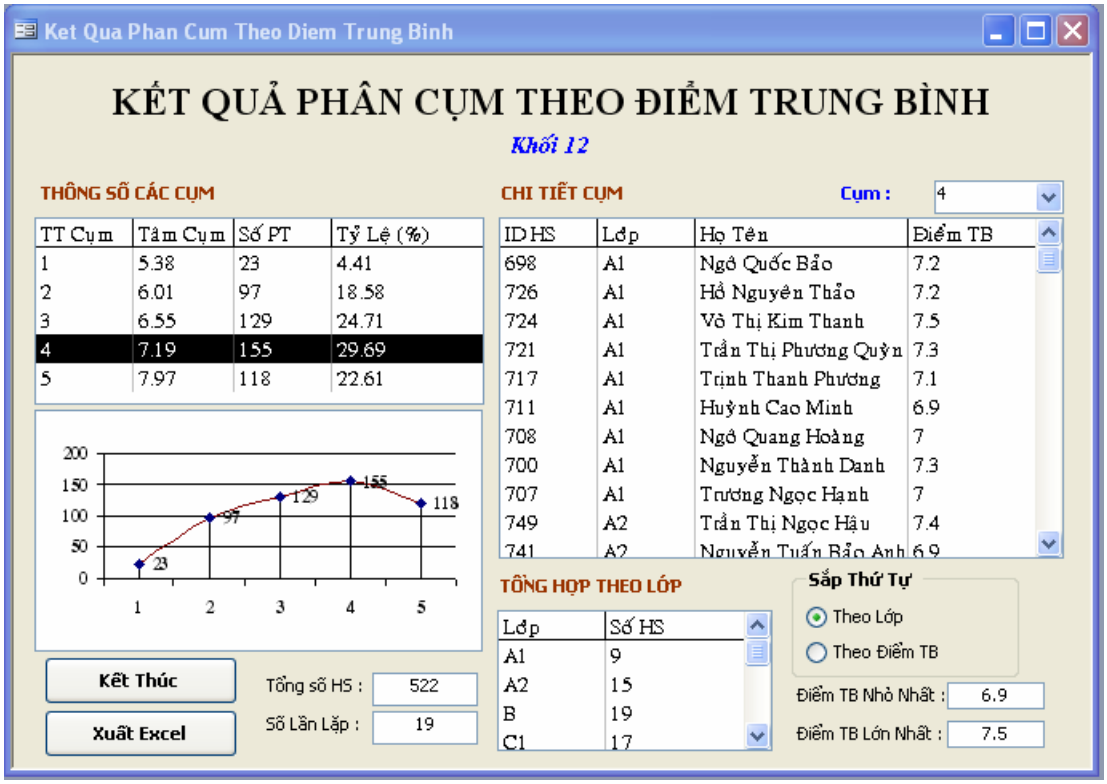
```
' Ghi chỉ số cụm của từng HS vào tập tin dữ liệu
Set dbf = CurrentDb
dbf.Execute "Delete * from kqcumhs_mon;"
Set db = dbf.OpenRecordset("kqcumhs_mon", dbOpenDynaset)
For i = 1 To SoMT
   db.AddNew
   db!idhs = AHS(i, 1)
   db!lop = AHS(i, 2)
   db!hoten = AHS(i, 3)
   tong = 0
   For m = 1 To SoMonHoc_
      tenfield = "mon" & m
      db.Fields(tenfield).Value = Round(AHS(i, 3 + m), 1)
```

```
        tong = tong + AHS(i, 3 + m)
    Next
    db!TTCum = AHS(i, 3 + SoMonHoc_ + 1)
    db!tongdiem = tong
    db.Update
Next
db.Close
```

Một số màn hình khi cho chạy dữ liệu đối với tập tin DiemChuVanAn2009.mdb

3.4.6. Một số chức năng thường sử dụng

1. Phân cụm theo điểm trung bình năm: Chọn lựa khối 12, 5 cụm

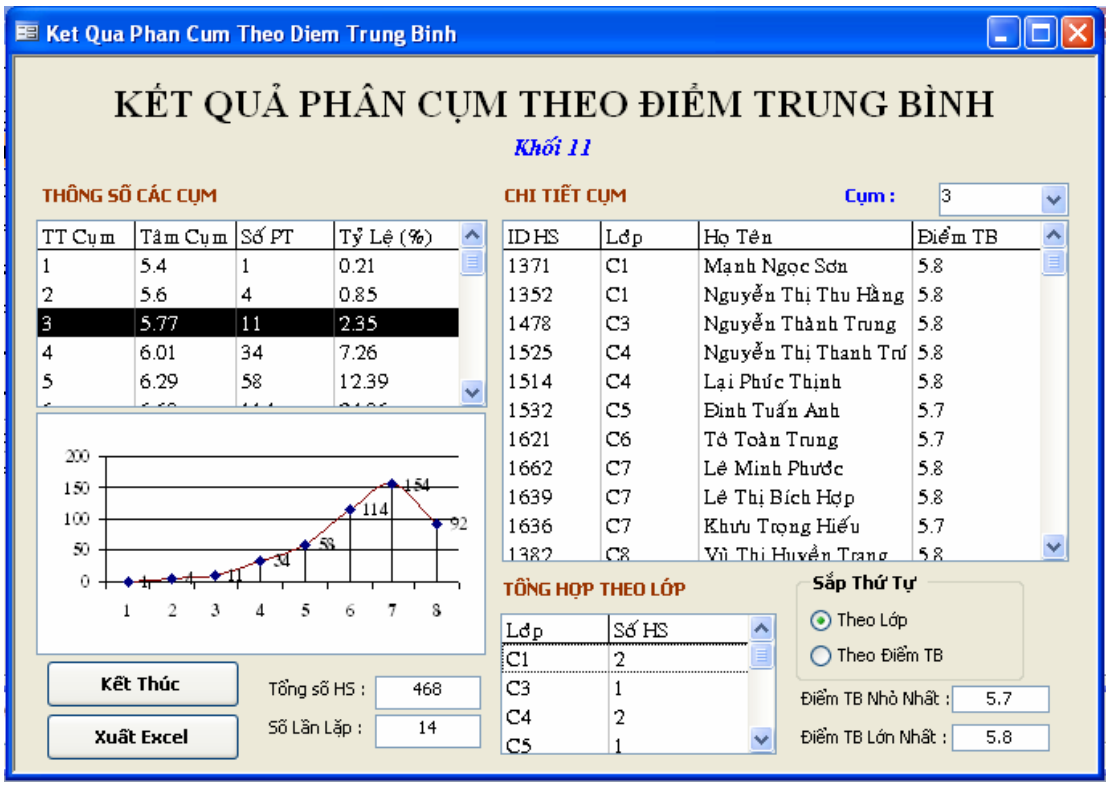


Hình 3.6 : Màn hình kết quả Chọn khối 12 và số cụm là 5

Trong màn hình này, người dùng có thể đưa ra được các phân tích với các nội dung (cụm 1 có số phần tử ít nhất, cụm 4 có số phần tử nhiều nhất) :

- Phần lớn học sinh học tập khá có điểm trung bình từ 6.9 đến 7.5. Đặc biệt có 118 học sinh giỏi có điểm trung bình từ 7.6 đến 8.9. Tập trung phần lớn học sinh ở lớp B, C1 và C4 cần tìm hiểu và nhân rộng hình thức học tập của các lớp này.
- Cá biệt có một số em (23 em) học yếu so với lớp có điểm trung bình từ 5.0 đến 5.6. Tập trung phần lớn học sinh ở lớp C2 và C6 cần rút kinh nghiệm công tác chủ nhiệm của các lớp này.
- Các học sinh yếu so với học sinh trong khối chỉ tập trung ở các lớp Ban C, cần xem lại số học sinh này có phải chọn nhầm Ban hay không?

2. Phân cụm theo điểm trung bình năm : Chọn lựa khối 11, 8 cụm



Hinh 3.7 : Màn hình kết quả Chọn khối 11 và số cụm là 8

Trong màn hình này, người dùng có thể đưa ra được các phân tích với các nội dung sau (cụm 1,2,3 có số phần tử ít , cụm 6,7,8 có số phần tử nhiều) :

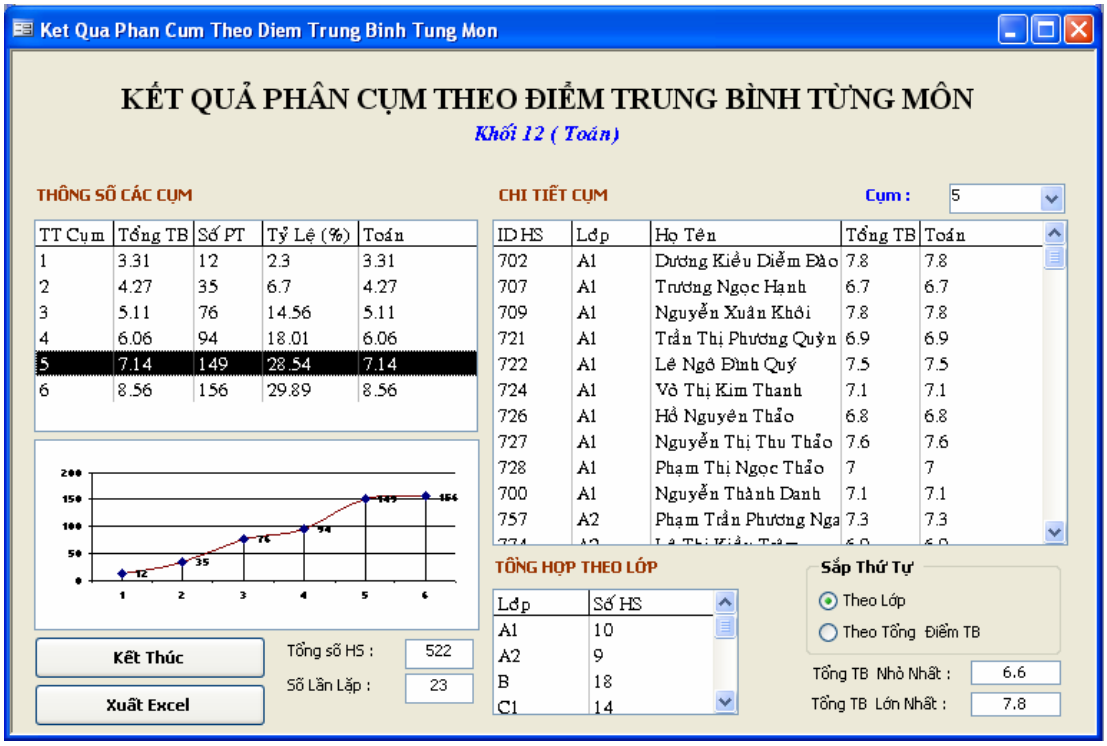
- Có sự tách biệt kết quả học tập của học sinh các lớp ban A và các lớp ban C, học sinh ban A tập trung ở cụm 5,6,7 và 8 còn học sinh ban C thì tập trung ở cụm 1,2,3 và 4.
- Phần lớn học sinh học tập khá có điểm trung bình từ 6.5 trở lên tập trung ở cụm 6,7 và 8. Đặc biệt số học sinh ban C có điểm trung bình càng cao có dấu hiệu giảm dần.
- Học sinh có điểm trung bình thấp ở cụm 1,2 và 3 lại tập trung ở ban C không có học sinh ban A (thậm trí ở cụm 4 chỉ có 1 học sinh ban A).

Từ các nội dung nhận xét trên Ban giám hiệu cần phải xem xét lại các vấn đề liên quan đến việc dạy và học ở ban C : từ đội ngũ giáo viên, chất lượng đầu vào của học sinh, công tác chủ nhiệm hay cơ sở vật chất, môi trường học tập cũng như việc chọn ban của học sinh.

3. Phân cụm theo điểm TB các môn học : Chọn khối 12; 6 cụm; Số nhóm phân tích là 1; chọn 1 môn Toán

Trong màn hình này, người dùng có thể đưa ra được các phân tích với các nội dung sau (số phần tử phần lớn tập trung ở cụm 4,5, và 6) :

- Theo số liệu phân tích khối lớp 12 học khá môn Toán thể hiện ở cụm 4, 5 và 6 với 399/522 học sinh với điểm bình quân môn Toán là 5.6 trở lên, tập trung ở lớp A1, lớp C1 và lớp B.
- Bên cạnh có những học sinh quá yếu môn Toán ở cụm 1 và 2 có điểm trung bình môn Toán từ 2.5 đến 4.6, những học sinh này đều tập trung ở những lớp Ban C



Hình 3.8 : Kết quả Chọn khối 12, số cụm là 8, phân tích 1 nhóm, môn Toán

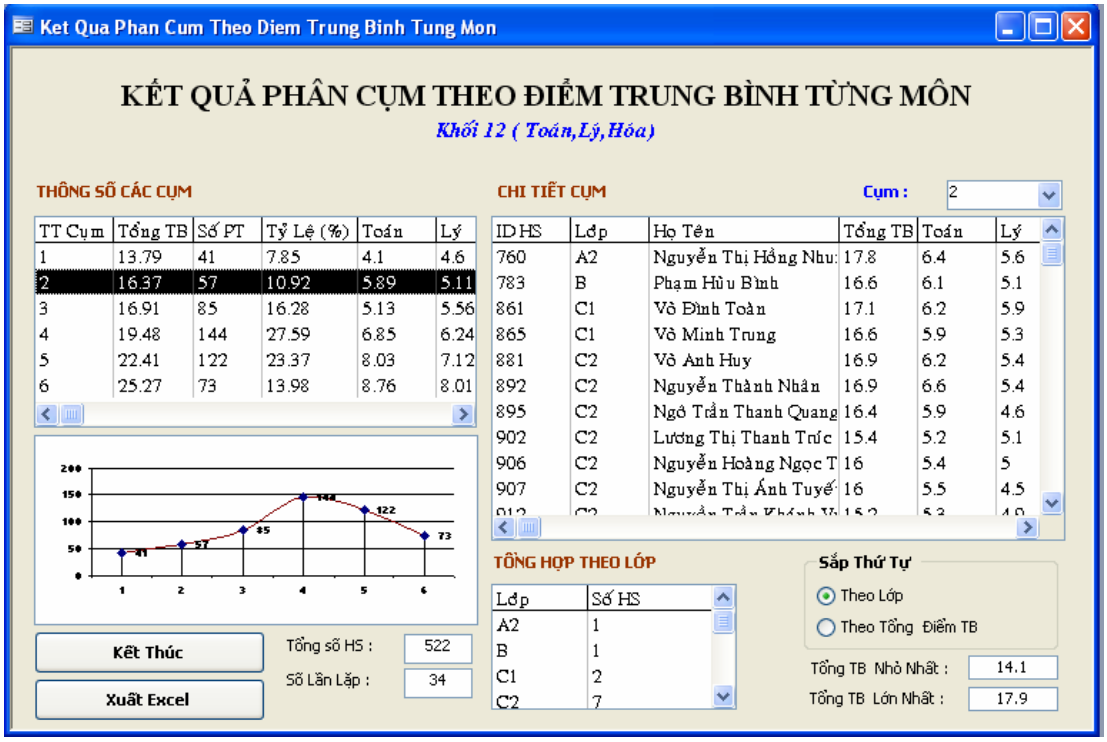
Từ các nội dung nhận xét trên Ban giám hiệu cần phải xem xét lại các vấn đề liên quan đến việc dạy và học môn Toán : các lớp C1 và lớp B không phải là lớp ban A nhưng vẫn có số học sinh giỏi toán chiếm khá nhiều, cần tìm hiểu giáo viên giảng dạy Toán ở 2 lớp này để có hình thức nhân rộng ra các lớp khác; cần có kế hoạch phụ đạo thêm môn Toán cho các lớp ban C vì phần lớn các lớp ban C đều có điểm kém về môn Toán.

4. Phân cụm theo điểm TB các môn học : Chọn khối 12; 6 cụm; Số nhóm phân tích là 1; chọn 3 môn Toán, Lý, Hóa

Trong màn hình này, người dùng có thể đưa ra được các phân tích với các nội dung sau (tập trung ở cụm 4 với 144/522 học sinh và cụm 5 với 122/522 học sinh) :

- Vì phân cụm dựa vào 3 môn Toán, Lý, Hóa từ đó thấy được sự vượt trội của ban A so với ban C, số học sinh ở cụm 4, 5, 6 có tổng điểm

trung bình 3 môn từ 18.0 điểm trở lên đều tập trung ở các lớp ban A. Nhưng đặc biệt vẫn có một số học sinh ban C ở lớp C1 (27 học sinh), C3 (35 học sinh) nằm trong các cụm này.



Hinh 3.9 : Màn hình kết quả Chọn khối 12, số cụm là 6, phân tích 1 nhóm, môn Toán, Lý và Hóa

- Ở cụm 1, 2 tổng điểm trung bình 3 môn từ 17.0 điểm trở xuống đều tập trung ở các lớp ban C. Tuy nhiên, cá biệt trong số này có 2 học sinh không phải ban C là Nguyễn Thị Hồng Nhung (lớp A2) và Phạm Hữu Bình (lớp B).

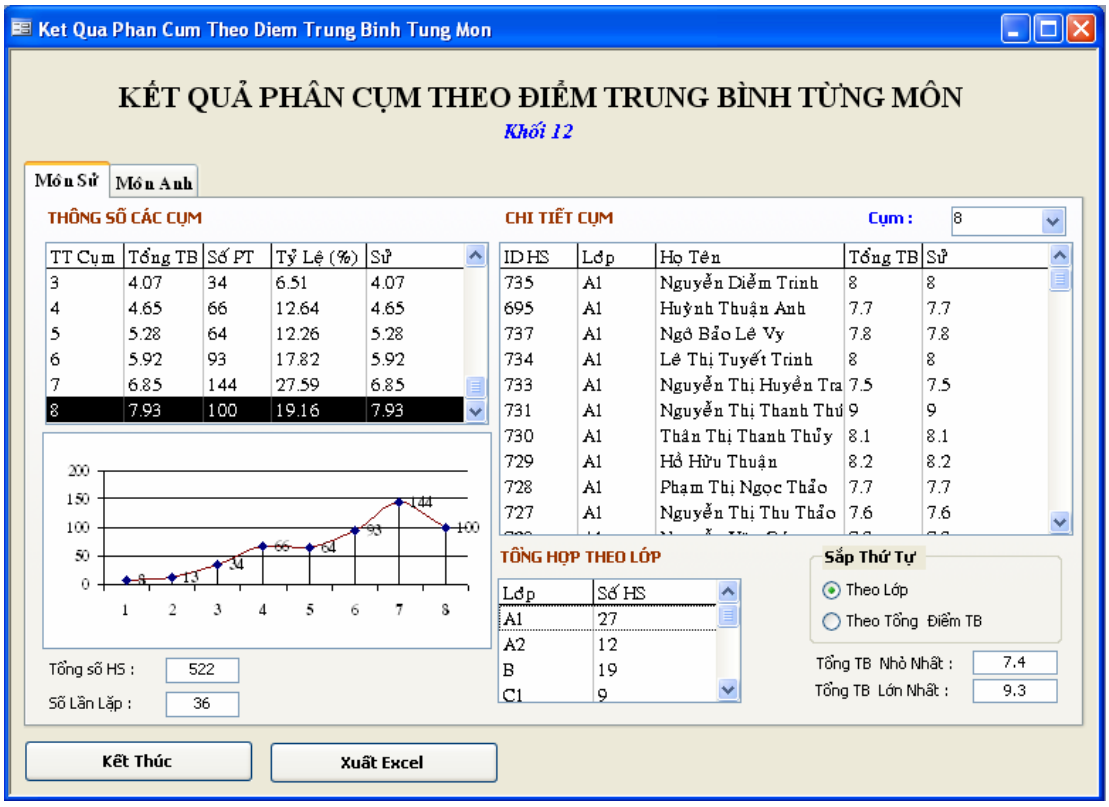
- Số học sinh có tổng điểm trung bình 3 môn từ 15.4 điểm trở lên chiếm trên 82% (424/522 học sinh) thể hiện sự vượt trội môn Toán, Lý, Hóa trong trường.

Từ các nhận xét trên Ban Giám hiệu cần xem lại 2 học sinh Nguyễn Thị Hồng Nhung (lớp A2) và Phạm Hữu Bình (lớp B) có chọn nhầm ban hay

không ?, một số học sinh trong lớp C1, C3 có dấu hiệu vượt trội về điểm môn Toán, Lý, Hóa có nên chuyển sang ban A hay không ?, cũng cần tìm hiểu giáo viên giảng dạy môn Toán, Lý, Hóa ở 2 lớp này là những giáo viên nào? Để có hình thức khuyến khích, nhân rộng kinh nghiệm giảng dạy.

5. Phân cụm theo điểm TB các môn học : Chọn khối 12; 6 cụm; Số nhóm phân tích là 2; chọn nhóm 1 : Sử và nhóm 2 : Ngoại ngữ. Trình bày 2 trang

- Trang môn Sử

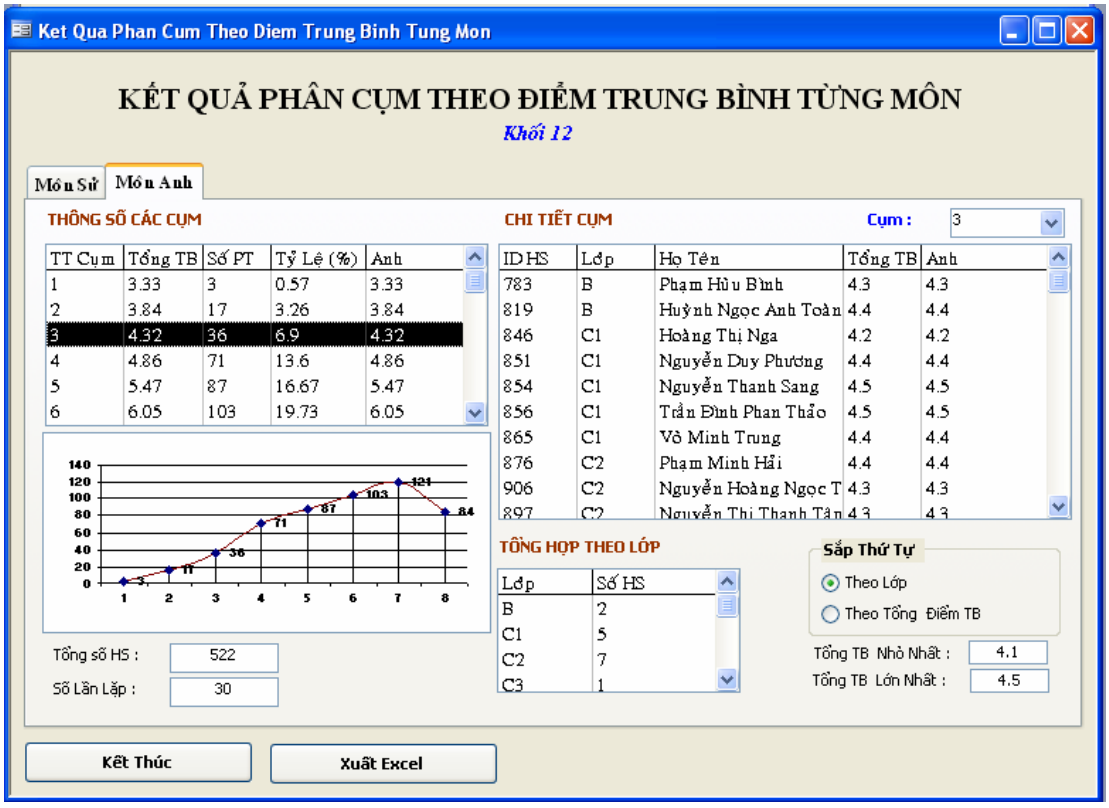


Hình 3.10 : Màn hình kết quả Môn Sử . Chọn khối 12, số cụm là 6, phân tích 2 nhóm, 2 môn Sử và Anh

Trong trang này, người dùng có thể đưa ra được các phân tích với các nội dung sau (tập trung ở cụm 6,7 và 8 với 337/522 học sinh) :

- Dựa trên môn Sử tuy nhiên số lượng có điểm từ 5.0 trở lên ở cụm 5, 6, 7, 8 phần lớn là học sinh ban A. Đặc biệt, cụm 8 có điểm trung bình từ 7.4 đến 9.3 lại có đến 58/100 học sinh không phải ban C (gồm lớp A1 27 học sinh, A2 12 học sinh và lớp B 19 học sinh).
- Ở cụm 1, 2, 3 và 4 là những cụm có điểm trung bình môn Sử từ 4.9 trở xuống, toàn bộ học sinh trong 4 cụm này đều là học sinh ban C (không có 1 học sinh ban A nào!). Đặc biệt lớp C2, C6 là 2 lớp góp mặt trong số lượng này khá nhiều (chiếm 38/121 học sinh).

  - Trang môn Anh

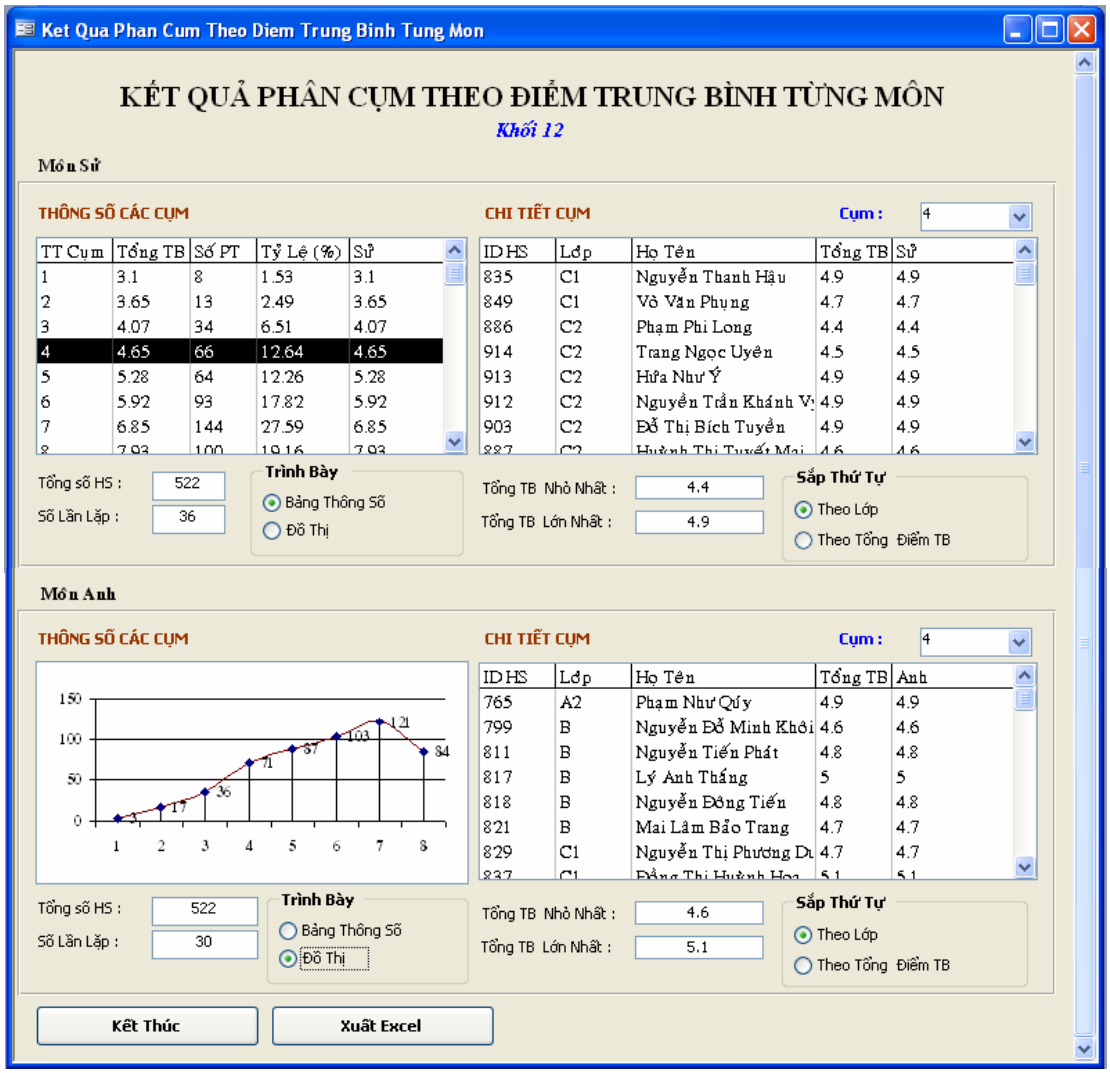


Hình 3.11 : Màn hình kết quả Môn Anh . Chọn khối 12, số cụm là 6, phân tích 2 nhóm, 2 môn Sử và Anh

Trong trang này, người dùng có thể đưa ra được các phân tích với các nội dung sau (tập trung ở cụm 5, 6, 7 và 8 với 405/522 học sinh) :

- Số lượng có điểm từ 5.2 trở lên ở cụm 5, 6, 7, 8 chiếm gần 80% học sinh cả khối, số lượng ban C trong số này chiếm đa số 286/405 học sinh. Đặc biệt, cụm 7 có điểm từ 6.5 đến 7.4 chỉ có 22/121 học sinh.

- Ở các cụm 1, 2, 3 và cụm 4 là những cụm có điểm trung bình môn Anh từ 5.1 trở xuống phần đông là học sinh ban C chỉ có một vài trường hợp nhỏ lẻ không phải ban C như : Nguyễn Hữu Bình, Huỳnh Ngọc Anh Toàn lớ B, Phạm Như Quý lớp A2.

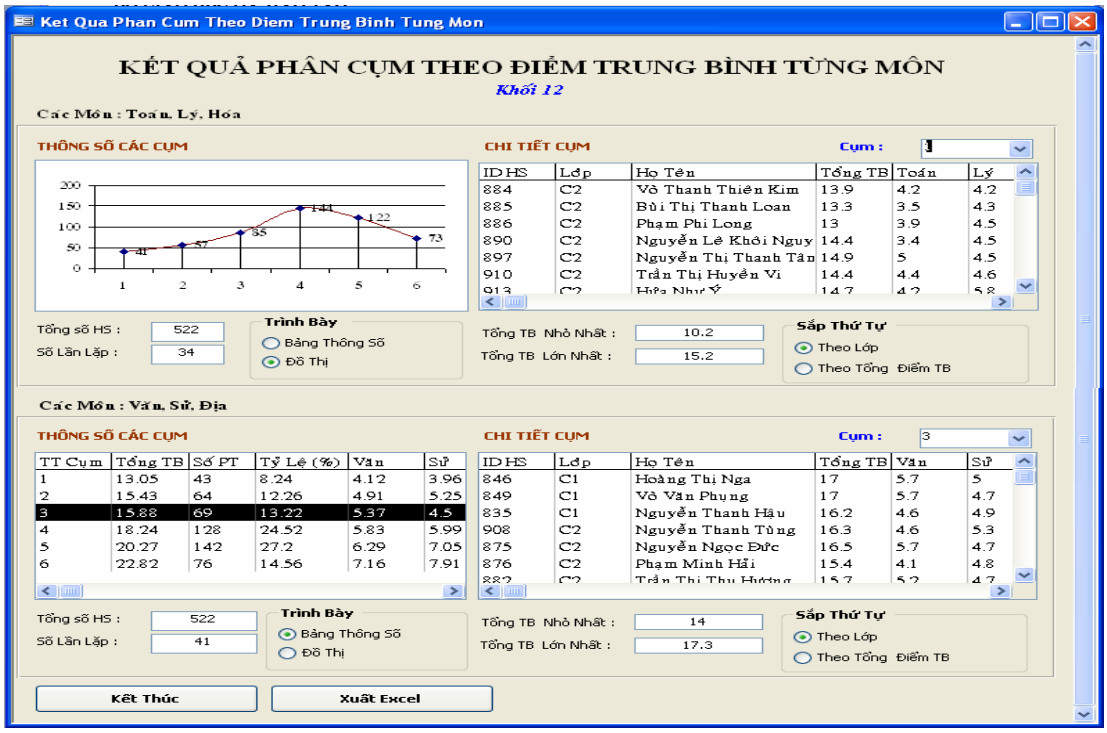


Hình 3.12 : Màn hình kết quả 2 môn cùng lúc . Chọn khối 12, số cụm là 6, phân tích 2 nhóm, 2 môn Sử và Anh

Qua 2 trang môn Sử và môn Anh vừa nêu trên Ban Giám hiệu cần xem lại : tình hình học tập ở ban C có sự phân cấp quá rõ ràng giữa 2 loại học sinh khá giỏi và yếu kém, nhằm có biện pháp hay kế hoạch phụ đạo các học sinh yếu kém; tình hình học môn Anh văn khả quan hơn môn Lịch sử nhưng cá biệt vẫn có một số lớp có học sinh yếu đều luôn cả 2 môn cần xem lại số học sinh này …Cần xem lại việc học tập của các em Nguyễn Hữu Bình, Huỳnh Ngọc Anh Toàn lớp B, Phạm Như Quý lớp A2 có cần chuyển lớp, chuyển ban hay không?

Tất nhiên, các sự phân tích, so sánh trên có thể dễ dàng hơn nếu chúng ta chọn thể hiện một trang màn hình khi phân tích (Hình 3.12).

6. Phân cụm theo điểm TB các môn học : Chọn khối 12; 6 cụm; Số nhóm phân tích là 2; chọn nhóm 1: Toán, Lý, Hóa và nhóm 2: Văn, Sử, Địa. Trình bày 1 trang màn hình



Hinh 3.13 : Màn hình kết quả 2 nhóm môn cùng lúc . Chọn khối 12, số cụm là 6, phân tích 2 nhóm, Toán, Lý, Hóa và Văn, Sử, Địa

Qua phân cụm kết quả so sánh 2 nhóm môn Toán, Lý, Hóa và Văn, Sử, Địa như trên Ban Giám hiệu nhà trường có thể : định hướng cho học sinh lựa chọn khối thi, trường thi trong kỳ thi tuyển sinh Đại học và Cao đẳng; có kế hoạch phụ đạo cho các em còn yếu các môn trong ban; nếu phân cụm trên thực hiện ở khối lớp 10 thì đó sẽ là cơ sở để cho việc xếp lớp; đây cũng có thể là tiền đề để có kế hoạch bồi dưỡng giáo viên.

Tất nhiên, trên đây mới chỉ là các đề xuất của tác giả trong việc đưa ra phân tích, nhận xét sau khi phân cụm dữ liệu với các ví dụ cụ thể. Trong thực tế đối với mỗi đơn vị trường, mỗi cá nhân … còn có nhiều yêu cầu phân cụm khác như : phân cụm để so sánh 6 môn thi tốt nghiệp ở học sinh 12; phân cụm để chọn lớp chọn, lớp năng khiếu; phân cụm để chọn các em cần phụ đạo; phân cụm để định hướng cho các em chọn ban dự thi Đại học v.v…

Để thuận tiện cho việc sử dụng dữ liệu dài lâu và cho các công việc khác nhau, trong các cửa sổ kết quả của chương trình đều có chức năng Xuất ra Excel để người dùng thuận tiện trong thao tác kết xuất dữ liệu ra ngoài.

3.5. Kết luận

Phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh dựa trên điểm trung bình môn học, điểm trung bình học kỳ … bước đầu ít nhiều đã giúp cho Ban giám hiệu nhà trường, các nhà quản lý giáo dục có được một cái nhìn nhiều chiều hơn, đa dạng hơn, nhiều góc cạnh hơn về điểm số của học sinh từ đó thu được một số kết quả như : việc phân lớp, lựa chọn học sinh giỏi để bồi dưỡng, phát hiện học sinh yếu kém để phụ đạo … cũng như đề ra kế hoạch giảng dạy, tăng giờ tăng tiết, định hướng nghề nghiệp cho học sinh qua việc chọn ban và khối thi Đại học v.v…

KẾT LUẬN

Trong quá trình tìm hiểu và hoàn thành luận văn tốt nghiệp với đề tài "Ứng dụng phân cụm dữ liệu trong việc phân tích, đánh giá kết quả học tập của học sinh", dù đã đạt được một số kết quả nhất định về kiến thức, về thực tế (chương trình Phân tích, đánh giá kết quả học tập của học sinh qua phân cụm dữ liệu đã được sử dụng ở nhiều Trường Trung học Phổ thông trong tỉnh), nhưng bản thân nhận thấy phân cụm trong khai phá dữ liệu vẫn là một lĩnh vực nghiên cứu còn quá rộng lớn và còn đầy triển vọng bao hàm nhiều phương pháp, kỹ thuật, nhiều hướng nghiên cứu, tiếp cận khác nhau.

Đề tài đã cố gắng tập trung tìm hiểu, nghiên cứu, trình bày được một số kỹ thuật và thuật toán phân cụm dữ liệu phổ biến, dựa trên các phương pháp đã có, cài đặt thử nghiệm thuật toán k-means vào chương trình.

Với những gì mà luận văn đã thực hiện và đạt được, hướng phát triển sau này của luận văn như sau:

Về thực tiễn : sẽ phát triển thành bài toán cấp độ sở với số dữ liệu lớn hơn, bao quát hơn, nhiều chọn lựa hơn như phân cụm dựa trên: các loại hình trường công lập ngoài công lập, các trường ở các vùng miền khác nhau, các trường nằm trên địa bàn theo đơn vị hành chính của tỉnh …

Về lý thuyết : tiếp tục nghiên cứu tiếp cách phương pháp, các cách tiếp cận mới về phân cụm dữ liệu như : phân cụm thống kê, phân cụm khái niệm, phân cụm mờ, phân cụm mạng KOHONEN …tìm kiếm, so sánh và chọn lựa thuật toán tối ưu nhất để giải quyết bài toán đã đưa ra.

Mặc dù đã cố gắng tập trung nghiên cứu và tham khảo nhiều tài liệu, bài báo, tạp chí khoa học trong và ngoài nước, nhưng do trình độ còn có nhiều giới hạn không thể tránh khỏi thiếu sót và hạn chế, rất mong được sự chỉ bảo đóng góp nhiều hơn nữa của các quý thầy cô giáo và các nhà khoa học…

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Nguyễn Hoàng Tú Anh Giáo trình "Khai thác dữ liệu và ứng dụng" 2009 (Đại học KHTN Tp Hồ Chí Minh)

[2]. An Hồng Sơn Luận văn thạc sĩ "Nghiên cứu một số phương pháp phân cụm mờ và ứng dụng" 2008 (Đại học Thái Nguyên)

[3]. Vũ Lan Phương "Nghiên cứu và cài đặt một số giải thuật phân cụm phân lớp" 2006 (Đại học Bách khoa Hà Nội)

Tiếng Anh

[4]. Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides" Nov 2001 http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html

[5]. Dr. Osmar R.Zaiane "Principles of knowledge discovery in databases" Fall 2001 (University of Alberta)

[6]. Patrick André Pantel "Clustering by Committee" Thesis Doctor of Philosophy, Spring 2003 (University of Alberta), 15 - 25p

[7]. Jiawei Han and Micheline Kamber "Data Mining Concepts and Techniques" 2007 Chapter 1 & Chapter 8 (Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada)

[8]. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html