Chương 5 : Phân cụm

Trịnh Anh Phúc [1]

[1] Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

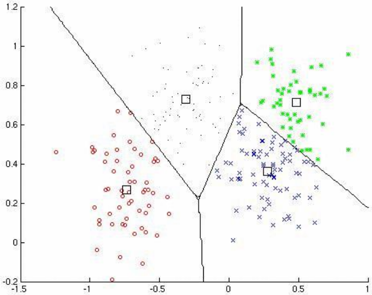
Ngày 14 tháng 11 năm 2014

Giới thiệu

Bài toán phân cụm

Định nghĩa của bài toán phân cụm

Tìm kiếm các nhóm gồm các đối tượng sao cho các đối tượng cùng một nhóm thì giống (liên quan) với nhau còn các đối tượng khác nhóm thì không giống (không liên quan) với nhau.



Bài toán phân cụm

Nhận xét

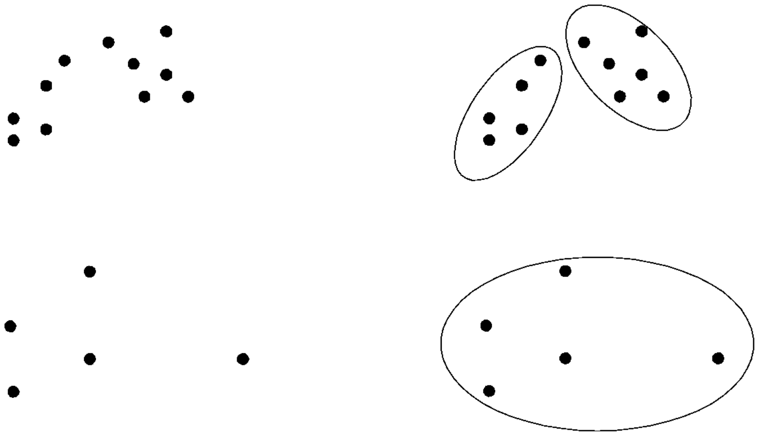
- Giống như mô hình k-hàng xóm gần nhất, khoảng cách thường dùng làm thước đo sự giống nhau giữa các đối tượng dữ liệu
- Số các nhóm (cụm) k cũng thường được xác định cảm tính
- Dữ liệu ở đây là dữ liệu không gán nhãn lớp nên thường không có cách đánh giá giải thuật phân cụm là tốt hay không
- Tuy nhiên, bài toán phân cụm có *ứng dụng rất nhiều* trong thực tế

Bài toán phân cụm

Phân loại các bài toán phân cụm

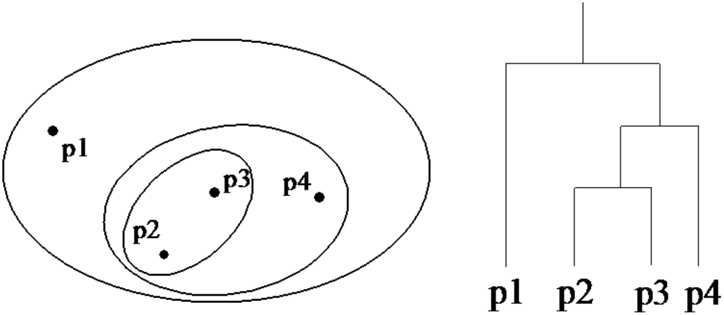
- Phân cụm theo nhóm (partitional clustering) dẫn suất trực tiếp từ định nghĩa
- Phân cụm theo cấu trúc cây (hierarchical clustering)
- Phân cụm theo mật độ (density clustering)

Bài toán phân cụm



Hình: Hình mô tả phân cụm theo nhóm. Các điểm dữ liệu được tô đen bên trái còn sau khi phân cụm được bao lại bởi ba hình e líp

Bài toán phân cụm



Hình: Hình mô tả phân cụm theo cấu trúc cây. Các điểm dữ liệu được tô đen bên trái còn sau khi phân cụm được biểu diễn bởi một cây phân cụm, mỗi mức ứng với một hình e líp bao tập con điểm dữ liệu. Cảng mức thấp, số điểm dữ liệu càng lớn.

Dành cho trả lời câu hỏi

Bài toán phân cụm

Ý tưởng

- mỗi cụm được gán với một trung tâm (centroid)
- mỗi điểm dữ liệu được gán với một cụm nếu nó gần trung tâm của cụm đó nhất
- số k các cụm cần được chỉ rõ
- cơ bản đây là giải thuật đơn giản

Bài toán phân cụm

Thuật toán k-Means

- **Đầu vào** : Tập các điểm dữ liệu biểu diễn các đối tượng $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$
- **Đầu ra** : Tập $k = \{\mathbf{c}_j, \ldots, \mathbf{c}_k\}$ trung tâm tương ứng mỗi cụm

Function k-Means($\mathbf{X}$)

1. Khởi tạo ngẫu nhiên $k = \{\mathbf{c}_j\}$ với $j = 1, \ldots, k \in \mathbf{X}$
2. **forall**($\mathbf{x}_i$ thuộc tập $\mathbf{X}$) **do** Xác định trung tâm gần nhất với $\mathbf{x}_i$ **endfor**
3. **forall** ($\mathbf{c}_j$ thuộc tập k) **do** Tính lại các giá trị $\mathbf{c}_j$ **endfor**
4. Lặp lại bước 2-3 đến khi tập k không đổi
5. **return** k

End

Bài toán phân cụm

Xét tập dữ liệu số baskball.arff

TT	tham gia/phút	chiều cao	thời gian chơi	tuổi	số điểm/phút
1	0.0888	201.0	36.02	28.0	0.5885
2	0.1399	198.0	39.32	30.0	0.8291
3	0.0747	198.0	38.8	26.0	0.4974
4	0.0983	191.0	40.71	30.0	0.5772
5	0.1276	196.0	38.4	28.0	0.5703
6	0.1671	201.0	34.1	31.0	0.5835
7	0.1906	193.0	36.2	30.0	0.5276
8	0.1061	191.0	36.75	27.0	0.5523

Khởi tạo với hai trung tâm $k = \{\mathbf{c}_1, \mathbf{c}_2\}$

- $\mathbf{c}_1 = \{0.0747, 198.0, 38.8, 26.0, 0.4974\} \Leftarrow$ bản ghi thứ 3
- $\mathbf{c}_2 = \{0.1906, 193.0, 36.2, 30.0, 0.5276\} \Leftarrow$ bản ghi thứ 7

Bài toán phân cụm

Vòng lặp 1

tham gia/phút	chiều cao	thời gian chơi	tuổi	số điểm/phút	cụm $\mathbf{c}_j$
0.0888	201.0	36.02	28.0	0.5885	2
0.1399	198.0	39.32	30.0	0.8291	2
0.0747	198.0	38.8	26.0	0.4974	1
0.0983	191.0	40.71	30.0	0.5772	1
0.1276	196.0	38.4	28.0	0.5703	1
0.1671	201.0	34.1	31.0	0.5835	2
0.1906	193.0	36.2	30.0	0.5276	2
0.1061	191.0	36.75	27.0	0.5523	1

Tính lại hai trung tâm $k = \{\mathbf{c}_1, \mathbf{c}_2\}$

- $\mathbf{c}_1 = \{0.1017, 194.00, 38.665, 27.75, 0.5493\}$
- $\mathbf{c}_2 = \{0.1466, 198.25, 36.41, 29.75, 0.6322\}$

Bài toán phân cụm

Vòng lặp 2

- $\mathbf{c}_1 = \{0.1317, 191.6667, 37.8867, 29.00, 0.5524\}$
- $\mathbf{c}_2 = \{0.1196, 198.80, 37.328, 28.60, 0.6138\}$

Vòng lặp 3

- $\mathbf{c}_1 = \{0.1022, 191.00, 38.73, 28.50, 0.5648\}$
- $\mathbf{c}_2 = \{0.1314, 197.8333, 37.14, 28.8333, 0.5994\}$

Vòng lặp 4

- $\mathbf{c}_1 = \{0.1061, 191.00, 36.75, 27.00, 0.5523\}$
- $\mathbf{c}_2 = \{0.1267, 196.8571, 37.65, 29.00, 0.5962\}$

Vòng lặp 5

- $\mathbf{c}_1 = \{0.1061, 191.00, 36.75, 27.00, 0.5523\}$
- $\mathbf{c}_2 = \{0.1267, 196.8571, 37.65, 29.00, 0.5962\}$

Kết thúc khi các trung tâm $k = \{\mathbf{c}_1, \mathbf{c}_2\}$không thay đổi

Bài toán phân cụm

Nhận xét

- Hoàn toàn sử dụng được dữ liệu dạng số
- Giải thuật phụ thuộc vào k trung tâm khởi tạo ban đầu
- Chỉ tìm được điểm tối ưu cục bộ
- Để xác định xem các trung tâm không thay đổi, ta có thể tính bằng trị tuyệt đối hiệu số chuẩn của các trung tâm giữa hai vòng lặp liên tiếp

Bài toán phân cụm

Ý tưởng của thuật toán EM

- Ta coi mỗi cụm là một phân bố có tham số
- Quá trình xác định cụm chính là xác định tham số của các phân bố này
- Giải thuật làm tăng tối đa kỳ vọng của dữ liệu biểu diễn các đối tượng được quan sát, trên tập các tham số tương ứng k phân bố

Bài toán phân cụm

Các thành phần của mô hình (Đọc thêm)

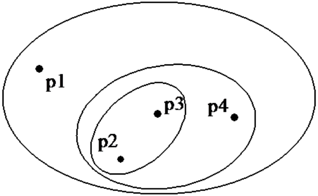
- Mô hình trộn
- Giải thuật Expectation Maximization (EM algorithm) dùng để ước lượng mô hình trộn

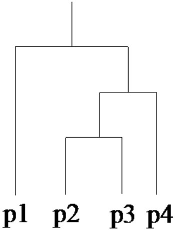
Dành cho trả lời câu hỏi

Bài toán phân cụm

Định nghĩa phân cụm theo cấu trúc cây

- Đầu ra là một tập các cụm-bao-chồng lên nhau (nested-clusters) tạo nên một cây
- Cây trên biểu diễn một chuỗi các phép nhập/chia cụm-bao-chồng





Bài toán phân cụm

Nhận xét về phân cụm theo cấu trúc cây

- Không cần khai báo số k cụm ban đầu như phân cụm theo nhóm. Chỉ cần cắt cây theo một mức nào đó thì ta có được một phân cụm theo nhóm có k nhóm tất cả
- Cách phân cụm này tương ứng phép phân loại trong xây dựng cây quyết định dưới-lên-trên

Bài toán phân cụm

Thuật toán phân cụm theo cấu trúc cây có yêu cầu sau

- **Đầu vào** : Tập các điểm dữ liệu $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$
- **Đầu ra** : Tập các cụm-bao-chồng $\mathbf{P}_1, \cdots, \mathbf{P}_n$ của $\mathbf{X}$ gồm lần lượt 1,2,....,n cụm sao cho $\sum_{i=1}^{n} cost(\mathbf{P}_i)$ là nhỏ nhất có thể

trong đó $cost(\mathbf{P}_i)$ được hiểu là chi phí phân chia bất kỳ tập điểm dữ liệu có kích thước i. Tùy thuộc vào cách tính chi phí $cost(\mathbf{P}_i)$, ta sẽ có được cây phân cụm khác nhau.

Bài toán phân cụm

Hai hướng phân cụm chính

- Bao dần (Agglomerative) :
  - Bắt đầu với mọi điểm dữ liệu đều làm trung tâm của n cụm
  - Mỗi bước, ta hợp hai cụm có chi phí - khoảng cách - gần nhau nhất thành một cụm mới tương ứng mức mới trên cây
- Chia tách (Divise) :
  - Bắt đầu với một cụm duy nhất gồm tất cả các điểm dữ liệu
  - Mỗi bước, ta chia một cụm thành hai cụm sao cho chi phí nhỏ nhất có thể. Quá trình kết thúc khi chỉ còn một điểm dữ liệu ở mỗi cụm

Bài toán phân cụm

Để thực hiện phân cụm theo cấu trúc cây, ta thường dùng ma trận khoảng cách. Ví dụ tập dữ liệu 8 điểm dữ liệu baskball.arff có ma trận như sau

	1	2	3	4	5	6	7	8
1	0.00	4.89	4.55	11.22	5.54	3.56	8.25	10.08
2	4.89	0.00	4.05	7.14	2.99	6.11	5.90	8.04
3	4.55	4.05	0.00	8.29	2.86	7.49	6.91	7.36
4	11.22	7.14	8.29	0.00	5.86	12.03	4.93	4.97
5	5.54	2.99	2.86	5.86	0.00	7.25	4.22	5.36
6	3.56	6.11	7.49	12.03	7.25	0.00	8.33	11.09
7	8.25	5.90	6.91	4.93	4.22	8.33	0.00	3.65
8	10.08	8.04	7.36	4.97	5.36	11.09	3.65	0.00

Bài toán phân cụm

Giải thuật phân cụm Agglomerative

- **Đầu vào** : Tập dữ liệu $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$
- **Đầu ra** : Cây biểu diễn phân cụm có n mức, mỗi lá biều diễn một điểm dữ liệu

Procedure AgglomerativeClustering($\mathbf{X}$)

1. Tạo ma trận khoảng cách
2. Cho mọi điểm dữ liệu $\mathbf{X}$ thành n cụm
3. **Repeat**
4. Hợp hai cụm gần nhau nhất
5. Cập nhật ma trận trọng số
6. **Until** chỉ còn lại một cụm

End

Bài toán phân cụm

Vòng lặp 1 : Hai điểm dữ liệu gần nhất là $d_{min} = 2.86 = (3, 5)$

	1	2	3	4	5	6	7	8
1	0.00	4.89	4.55	11.22	5.54	3.56	8.25	10.08
2	4.89	0.00	4.05	7.14	2.99	6.11	5.90	8.04
3	4.55	4.05	0.00	8.29	2.86	7.49	6.91	7.36
4	11.22	7.14	8.29	0.00	5.86	12.03	4.93	4.97
5	5.54	2.99	2.86	5.86	0.00	7.25	4.22	5.36
6	3.56	6.11	7.49	12.03	7.25	0.00	8.33	11.09
7	8.25	5.90	6.91	4.93	4.22	8.33	0.00	3.65
8	10.08	8.04	7.36	4.97	5.36	11.09	3.65	0.00

- Hợp hai điểm dữ liệu $\mathbf{x}_3$ và $\mathbf{x}_5$ thành một cụm mới
- Tính lại trung tâm của cụm mới và khoảng cách với các điểm còn lại $\Rightarrow$ cập nhật ma trận tại các ô tô xanh

Bài toán phân cụm

Vòng lặp 2 : Điểm dữ liệu gần nhất là $d_{min} = 3.26 = (2, \{3, 5\})$

	1	2	3	4	5	6	7	8
1	0.00	4.89	4.86	11.22	4.86	3.56	8.25	10.08
2	4.89	0.00	3.26	7.14	3.26	6.11	5.90	8.04
3	4.86	3.26	0.00	7.03	0.00	7.23	5.55	6.28
4	11.22	7.14	7.03	0.00	7.03	12.03	4.93	4.97
5	4.86	3.26	0.00	7.03	0.00	7.23	5.55	6.28
6	3.56	6.11	7.23	12.03	7.23	0.00	8.33	11.09
7	8.25	5.90	5.55	4.93	5.55	8.33	0.00	3.65
8	0.08	8.04	6.28	4.97	6.28	11.09	3.65	0.00

- Hợp điểm dữ liệu $\mathbf{x}_2$ với cụm cũ gồm $\mathbf{x}_3$,$\mathbf{x}_5$ thành một cụm mới
- Tính lại trung tâm của cụm mới và khoảng cách với các điểm còn lại $\Rightarrow$ cập nhật ma trận tại các ô tô xanh

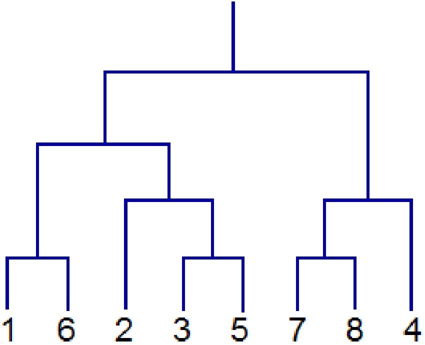
Bài toán phân cụm

- Vòng lặp 3 : Điểm dữ liệu gần nhất là $d_{min} = 3.56 = (1, 6)$
- Vòng lặp 4 : Điểm dữ liệu gần nhất là $d_{min} = 3.65 = (7, 8)$
- Vòng lặp 5 : Điểm dữ liệu gần nhất là $d_{min} = 4.6 = (4, \{7, 8\})$
- Vòng lặp 6 : Điểm dữ liệu gần nhất là $d_{min} = 5.34 = (\{1, 6\}, \{2, 3, 5\})$
- Vòng lặp 7 : Điểm dữ liệu gần nhất là $d_{min} = 7.91 = (\{1, 6, 2, 3, 5\}, \{7, 8, 4\})$

Cuối cùng ta có được tất cả các điểm dữ liệu trong một cụm

Bài toán phân cụm

Vậy ta có cây biểu diễn phân cụm theo giải thuật Agglomerative Clustering



Bài toán phân cụm

Khoảng cách giữa các cụm qui định hình thái của cây phân cụm

- Khoảng cách **single-link** giữa hai cụm C_i và C_j được định nghĩa như sau

$$D_{sl} = (C_i, C_j) = \min_{x,y}\{d(x,y) | x \in C_i, y \in C_j\}$$

- Khoảng cách **complete-link** giữa hai cụm C_i và C_j được định nghĩa như sau

$$D_{cl} = (C_i, C_j) = \max_{x,y}\{d(x,y) | x \in C_i, y \in C_j\}$$

- Khoảng cách **centroid** giữa hai cụm C_i và C_j được định nghĩa như sau

$$D_{centroids} = (C_i, C_j) = d(\mathbf{c}_j, \mathbf{c}_i)$$

với $\mathbf{c}_j, \mathbf{c}_i$ là trung tâm hai cụm, i.e. áp dụng ví dụ baskball.arff

Bài toán phân cụm

Khoảng cách Ward

Khoảng cách **Ward** giữa hai cụm C_i và C_j là hiệu số giữa tổng bình phương khoảng cách của các điểm dữ liệu với hai trung tậm $\mathbf{c}_j$ và $\mathbf{c}_i$ với $\mathbf{c}_{ij}$ là trung tâm hợp của hai cụm C_{ij}

$$D_W(C_i, C_j) = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)^2 + \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{c}_j)^2 - \sum_{\mathbf{x} \in C_{ij}} (\mathbf{x} - \mathbf{c}_{ij})^2$$

trong đó

- $\mathbf{c}_i$ là trung tâm của C_i
- $\mathbf{c}_j$ là trung tâm của C_j
- $\mathbf{c}_{ij}$ là trung tâm của C_{ij}

Bài toán phân cụm

Nhận xét về khoảng cách Ward dùng để phân cụm

- Khá giống với khoảng cách centroid
- Ít bị ảnh hưởng bởi nhiễu và dữ liệu ngoại lai
- Hướng đến trung tâm của toàn bộ dữ liệu
- Theo cấu trúc cây tương tự giải thuật k-means, có thể dùng để khởi tạo k-means

Bài toán phân cụm

Nhận xét chung về giải thuật Agglomerative Clustering

Với tập dữ liệu ban đầu $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, ta cần

- $O(n^2)$ là bộ nhớ chứa ma trận khoảng cách
- $O(n^3)$ trong hầu hết các trường hợp, gồm
  - $n-1$ bước để tìm và cập nhật ma trận trong số kích thước n
  - Có thể giảm xuống $O(n^2 \log n)$ bởi lựa chọn cấu trúc dữ liệu thích hợp cho ma trận

Dành cho trả lời câu hỏi

Bài toán phân cụm

Giải thuật phân cụm Devisive

- **Đầu vào** : Tập dữ liệu $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$
- **Đầu ra** : Cây biểu diễn phân cụm có n mức, mỗi lá biều diễn một điểm dữ liệu

Function DivisiveClustering($\mathbf{X}$,r)

1. Tách cụm $\mathbf{X}$ hai cụm xa nhau nhất $\mathbf{X}_i$ và $\mathbf{X}_j$ với $\mathbf{X} = \mathbf{X}_i \cup \mathbf{X}_j$
2. DivisiveClustering($\mathbf{X}_i$, r.left) // Gọi đệ qui
3. DivisiveClustering($\mathbf{X}_j$, r.right) // Gọi đệ qui
4. **return** r

End

Lần gọi đầu tiên DivisiveClustering($\mathbf{X}$,r) với $\mathbf{X}$ tập toàn bộ dữ liệu còn r là gốc cây

Bài toán phân cụm

Nhận xét về giải thuật DivisiveClustering

- Ít được dùng so với Agglomerative Clustering
- Có tới $O(2^n)$ phép chia một cụm $\mathbf{X}$ kích thước n thành hai cụm con $\mathbf{X}_i$ và $\mathbf{X}_j$
- Được gọi đệ qui sau mỗi lần chia
- Độ phức tập tính toán cao

Chương trình Weka

Yêu cầu

- Tải tập dữ liệu baskball.arff
- Chạy giải thuật SimpleKmeans và HierarchicalCluster
- Giải thích các tham số của các giải thuật