CƠ SỞ TOÁN HỌC CỦA PHÂN LỚP

Lê Đức Vĩnh

Đặt vấn đề

- Phương pháp phân lớp: Nếu ta coi tập A khác rỗng là một thực thể và tìm cách phân chia nó thành một số thực thể nhỏ
- Phương pháp ghép lớp: Nếu coi mỗi phần tử của tập A là một tập con của A và tìm cách ghép các tập con này lại với nhau thành một số tập.

- Bài toán phân lớp, ghép lớp được tất cả các ngành khoa học, từ khoa học tự nhiên, khoa học kỹ thuật đến khoa học xã hội đều quan tâm. Với mỗi tập gồm n phần tử, mỗi một chuyên ngành có yêu cầu đòi hỏi việc phân lớp ghép lớp khác nhau.
- Các phương pháp phân lớp theo truyền thống thường chỉ chú trọng tới:
  - Một rất số ít các đặc tính được họ coi là quan trọng,
  - Dựa trên kiến thức khoa học
  - Kinh nghiệm chuyên môn của bản thân

- Phân lớp chỉ dựa trên một số rất ít các đặc tính sẽ có những khiếm khuyết: mang tính chủ quan, thiếu toàn diện, mất đi tính khách quan và cách nhìn toàn diện mà bất cứ ngành khoa học nào cũng luôn đòi hỏi.
- Bài trình bày sẽ đưa ra một phương pháp tiếp cận mới đối với bài toán phân lớp và ghép lớp.

Ý tưởng của phương pháp

- Mỗi cá thể trong một đám đông các cá thể gồm có m đặc tính X_1; X_2;…;X_m

-Mỗi cá thể trong đám đông cần phân loại, tương ứng với một bộ m số $\left(X_1, X_2, ..., X_m\right) \in \mathbb{R}^m$

-Trong $\mathbb{R}^m$ ta xây dựng một độ đo thích hợp đo sự " gần gũi" giữa các cặp cá thể (các cặp phần tử) trong đám đông. Dựa vào số đo về sự "gần gũi" ta đưa ra phương pháp phân lớp, ghép lớp thích hợp.

1. *Khoảng cách, Metric, siêu Metric* :

Cho $A \neq \varnothing$, ánh xạ d gọi là hàm khoảng cách trong A nếu nó thỏa mãn:

1) $\forall a,b \in A,\ d(a,b) \geq 0\ ;\ d(a,b) = 0 \Leftrightarrow a = b$

2) $\forall a,b \in A\ ,\ d(a,b) = d(b,a)$

Hàm khoảng cách trong A gọi là metric nếu:

$$\forall a,b,c \in A\ ,\ d(a,c) \leq d(a,b) + d(b,c)$$

Hàm khoảng cách trong A gọi là siêu metric nếu:

$$\forall a,b,c \in A\ ,\ d(a,b) \leq \sup\left\{d(a,c); d(b,c)\right\}$$

Định lý 1: Một siêu metric là metric

2. Tập các cá thể, ma trận số liệu, trọng tâm đám đông

$A = \{a_1; a_2; ....; a_n\}$ mỗi cá thể thuộc A có m đặc tính định lượng X_1; X_2;…; X_m.

$$a_i = (x_{i1}; x_{i2}; ...; x_{im})$$

1) Ma trận $X = [x_{ij}]_{n\times m}$ gọi là ma trận các số liệu

2) $\overline{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ giá trị trung bình của biến X_k

3) $\overline{x} = (\overline{x_1}; \overline{x_2}; .....; \overline{x_m})$: trọng tâm của tập A

4) $Y = [y_{ik}]_{n\times m}$: ma trận quy tâm.

$$y_{ik} = x_{ik} - \overline{x_k}$$

3. Một số khoảng cách trong $\mathbb{R}^m$

- Squared Euclidean distance $d_1(x,y) = \sum_{i=1}^{m} (x_i - y_i)^2$
- Euclidean distance $d_2(x,y) = \left[\sum_{i=1}^{m} (x_i - y_i)^2 \right]^{\frac{1}{2}}$
- City-block distance hoặc Manhattan distance $d_3(x,y) = \sum_{i=1}^{m} |x_i - y_i|$
- Chebychev distance $d_4(x,y) = Maximum |x_i - y_i|$
- Power distance $d_5(x,y) = \left[(x_i - y_i)^p \right]^{\frac{1}{r}}$

- Khoảng cách Mahalanobis

$$d_6(a_i;a_t) = \left[(a_i - a_t)S^{-1}(a_i - a_t)^t\right]^{\frac{1}{2}}$$

- Khoảng cách có trọng số: gán cho X_j trọng số f_j: $\sum_{j=1}^{m} f_j = 1$

$$d_7(a_i, a_t) = \left[(a_i - a_t)F(a_i - a_t)^t\right]^{\frac{1}{2}} \qquad F = \begin{bmatrix} f_1 & 0.....0.....0 \\ 0 & f_2.....0....0 \\ & ..... \\ 0.....0.... & f_m \end{bmatrix}$$

- **Định lý:** Khoảng cách Mahalanobis là bất biến đối với phép biến đổi tuyến tính không suy biến trong $\mathbb{R}^m$

4. Khoảng cách giữa các tập con rời nhau

$C(A) = \{C_1; C_2; ...; C_l\}$ là một phân hoạch của A

$C_i \in C(A)$ là một lớp của tập hợp A

$\overline{c_i}$; $\overline{c_k}$ lần lượt là trọng tâm của lớp C_i và C_k

Một số hàm khoảng cách giữa các lớp này:

1) $d_1(C_i, C_k) = \min d(a,b)$, $a \in C_i; b \in C_k$ Single distance

2) $d_2(C_i, C_k) = d(\overline{c_i}, \overline{c_k})$

3) $d_3(C_i, C_k) = \sqrt{\frac{n_i n_k}{n_i + n_k}} d(\overline{c_i}, \overline{c_k})$

4) Khoảng cách Mahalanobis $d_4(C_i, C_k) = \left[(\overline{c_i}, \overline{c_k}) S^{-1} (\overline{c_i}, \overline{c_k})^t\right]^{\frac{1}{2}}$

S là ma trận Covariance của ma trận dữ liệu X

Một quan hệ hai ngôi H trong A là một quan hệ tiền thứ tự nếu thỏa mãn:

- Tính phản xạ $\forall a \in A, H(a,a)$
- Tính bắc cầu $\forall a,b,c \in A\ ,\ H(a,b)\&H(b,c) \Rightarrow H(a,c)$

Xét quan hệ H_T trong A sinh bởi quan hệ tiền thứ tự H: $\forall a,b \in A\ ,\ H_T(a,b) \Leftrightarrow H(a,b)\&H(b,a)$

- H_T là một quan hệ tương đương trong A

Giả sử H là một quan hệ hai ngôi trong $A \times A$ thỏa mãn:

1) $\forall a,b,c \in A$, $H[(a,a),(b,c)]$

2) $\forall a,b \in A$, $H[(a,b);(b,a)]$

3) $\forall a,b,c \in A$, $H[(b,c),(a,a)] \Rightarrow b = c$

Giả sử d là hàm khoảng cách trong A, ta xây dựng quan hệ hai ngôi H trong $A \times A$ dựa trên hàm khoảng cách d như sau:

$$\forall a,b,c,d \in A \text{ , } H[(a,b),(c,d)] \Leftrightarrow d(a,b) \le d(c,d)$$

Thì H là một quan hệ tiền thứ tự trong $A \times A$ nên sinh ra một quan hệ tương đương H_T

Quan hệ H xác định trong $C(A)\times C(A)$:

$$\forall C_1,C_2,C_3,C_4 \in C(A),\ H\left[(C_1,C_2),(C_3,C_4)\right] \Leftrightarrow d(C_1;C_2) \le d(C_3,C_4)$$

là một quan hệ tiền thứ tự trong $C(A)\times C(A)$

H_C là quan hệ tương đương sinh ra bởi hệ H

Xây dựng quan hệ H_C

$$\forall C_1,C_2,C_3 \in C(A),\ H_C(C_1,C_2) \Leftrightarrow H_T\left[(C_1,C_2),(C_2,C_3)\right]$$

Khi đó H_C là một quan hệ tương đương trong C(A)

H_C tạo ra các lớp tương đương trong C(A).

- Nếu $C(A) = \{\{a_1\}; \{a_2\} ..... \{a_n\}\} \equiv A$ thì với hàm khoảng cách d trong A, ta có thể tạo ra một phân hoạch chia A ra làm các lớp tương đương.
- Những kết quả nêu trên vừa là cơ sở toán học vừa chỉ ra cách phân lớp một tập đã cho theo khoảng cách. Phần sau đây sẽ giới thiệu các bước lập cây phân loại theo khoảng cách dựa vào các kết quả vừa trình bày.

Cây phân loại

- *Bước khởi đầu:* Lập hàm khoảng cách trong A, sau đó tính khoảng cách của các cặp phần tử.
- Tiếp theo ta lập hàm khoảng cách giữa các tập con rời nhau của A.
- Việc tạo lập cây phân loại được chia làm nhiều mức, ở mỗi mức các tập con này được ghép vào các lớp rời nhau, các lớp này tạo thành một phân hoạch của tập hợp A ở mức đang xét. Mức cuối cùng là mức mà phân hoạch thu được chỉ gồm một lớp đó là tập hợp A.

Khi ghép lớp phải tuân theo nguyên tắc sau

- Các lớp “gần” nhau sẽ được ghép lại với nhau ở những mức trước, các lớp “xa” nhau sẽ được ghép lại với nhau ở những mức sau.
- Mỗi lớp ở mức k (k>=1) chỉ gồm nhiều nhất 2 lớp ở mức k – 1
- Các cặp lớp được ghép với nhau ở mức k phải có cùng số đo sự “gần gũi” (khoảng cách) giữa các cặp lớp này.
- Nếu sau mức k – 1 có một lớp nào đó có cùng số đo về sự” gần gũi” với hai lớp khác thì lớp này chỉ được ghép với một trong hai lớp trên. Các lớp còn lại được giữ nguyên hoặc ghép với một lớp khác nếu chúng thỏa mãn nguyên tắc 3

Kết luận: những vấn đề được trình bày trong báo cáo trên đưa ra một cách tiếp cận mới của toán học đối với bài toán phân lớp và ghép lớp. Các nhà chuyên môn ở các lĩnh vực khác nhau có thể tìm thấy ở đây một phương pháp phân lớp mới với cách nhìn toàn diện và khách quan. Tuy vậy, nó chỉ là một phương pháp mới bổ sung cho phương pháp phân lớp truyền thống chứ không phải phương pháp tối ưu có thể thay thế các phương pháp phân lớp khác.

CẢM ƠN CÁC THẦY/CÔ ĐÃ LẮNG NGHE!