

# RAZVRŠČANJE Z DOMINANTNIMI MNOŽICAMI

Grupiranje (ang. Clustering) je postopek razvrščanja predmetov znotraj razreda v podrazrede (cluster) tako da so si predmeti znotraj istega podrazreda bolj podobni med sabo kot so si podobni z elementi iz ostalih podrazredov.

Problem združevanja lahko opišemo z uteženim grafom, ki ga definiramo kot trojico  $G=(V,E,w)$ , kjer je  $V=\{1,\dots,n\}$  končna množica vozlišč,  $E\subseteq V\times V$  množica usmerjenih povezav in  $\omega : E \rightarrow \mathbb{R}$  funkcija, ki vsakemu vozlišču dodeli neko vrednost (težo). Vozlišča grafa  $G$  ustrezajo predmetom, ki jih je potrebno združevati.

Povezave predstavljajo kateri predmeti so med seboj povezani, utežene povezave pa odražajo podobnosti med povezanimi predmeti. Poleg tega matrika  $A_{ij} = \omega(i,j)$  za vse  $i,j \in V$  predstavlja podobnost med vozlišči. Imenujemo jo matrika podobnosti.

Osnovni lastnosti, ki morata zadostovati gruči sta

- Notranja homogenost: elementi, ki pripadajo gruči si morajo biti med seboj podobni
- Maksimalnost: gruče ne moremo dodatno razširiti z uvedbo zunanjih elementov

## DEFINICIJE

Naj graf  $G$  predstavlja primer združevanja množic in naj bo  $C \subseteq V$  neprazna podmnožica. Povprečna utežena vhodna stopnja glede na  $C$  je definirana kot

$$awindeg_C(i) = \frac{1}{|C|} \sum_{j \in C} A_{ij}$$

kjer  $|C|$  predstavlja velikost množice  $C$ . Za  $j \in C$  definiramo

$$\phi_C(i,j) = A_{ij} - awindeg_C(j)$$

Funkcija  $\phi_C(i,j)$  je mera relativne podobnosti elementa  $i$  z elementom  $j$  glede na povprečno povezanost elementa  $i$  z elementi iz  $C$ .

Težo predmeta  $i$  glede na množico  $C$  definiramo kot

$$W_C(i) = \begin{cases} 1 & ; \text{ če } |C| = 1 \\ \sum_{j \in C \setminus \{i\}} \phi_C(i,j) W_{C \setminus \{i\}}(j); & \text{ sicer} \end{cases}$$

Vrednosti  $W_C(i)$  nam pove koliko podpore prejme element  $i$  od elementov  $C \setminus \{i\}$  glede na skupno podobnost z elementi iz  $C \setminus \{i\}$ . Pozitivne vrednosti nam povedo da je  $i$  močno koleriran z  $C \setminus \{i\}$ .

Skupna teža množice  $C$  pa je definirana z

$$W(C) = \sum_{i \in C} W_C(i)$$

### Definicija 1 (dominantna množica)

Neprazni množici  $C \subseteq V$  za katero je  $W(T) > 0$  za vsako neprazno množico  $T \subseteq C$  pravimo dominantna množica, če velja:

1.  $W_C(i) > 0$  za vse  $i \in C$
2.  $W_{C \cup \{i\}}(i) < 0$  za vse  $i \notin C$

### POVEZAVA S TEORIJO OPTIMIZACIJE

Če se omejimo na simetrične povezanosti, torej  $A$  je simetrična matrika, potem lahko dominantno množico zapišemo kot rešitev naslednjega standardnega kvadratičnega programa

$$\begin{aligned} \max f(x) &= x^T A x \\ \text{p.p. } x &\in \Delta \subset \mathbb{R}^n \end{aligned}$$

Kjer je  $\Delta = \{x \in \mathbb{R}^n : \sum_{j \in V} x_j = 1 \text{ in } x_j \geq 0 \text{ za vsak } j \in V\}$  standardni simpleks iz  $\mathbb{R}^n$ .

Pravimo, da je  $x$  rešitev zgornjega problema če obstaja soseščina  $x$ -a  $U \subseteq \Delta$  za katero je  $f(x) > f(z)$  za vsak  $z \in U \setminus \{x\}$ . Podpora  $\sigma(x)$  za  $x \in \Delta$  je definirana kot indeksna množica pozitivnih komponenta vektorja  $x$ , torej  $\sigma(x) = \{i \in V : x_i > 0\}$ .

### Definicija2 (otežen vektor)

Za neprazno podmnožico  $C$  množice  $V$  lahko definiramo otežen vektor  $x^C \in \Delta$  če ima množica  $C$  pozitivno skupno težo  $W(C)$ . V tem primeru je

$$x_i^C = \begin{cases} \frac{W_C(i)}{W(C)} & \text{če } i \in C \\ 0 & \text{sicer} \end{cases}$$

Za dominantno množico lahko torej vedno definiramo otežen vektor.

### Izrek1

Če je  $C$  dominantna množica  $A$ , potem je njen otežen vektor  $x^C$  rešitev zgornjega problema. Obratno, če je  $x^*$  rešitev zgornjega problema potem je njegova podpora  $\sigma = \sigma(x^*)$  dominantna množica od  $A$  pri pogoju da je  $W_{\sigma \cup \{i\}}(i) \neq 0$  za vse  $i \notin \sigma$ .

### POVEZAVA S TEORIJO GRAFOV

Naj bo  $G=(V,E)$  neusmerjen graf, kjer je  $V=\{1,2,...,n\}$  množica vozlišč in  $E \subseteq V \times V$  množica povezav v grafu. Dve vozlišči  $u, v \in V$  sta sosednju če  $(u, v) \in E$ . Podmnožici vozlišč  $C \subseteq V$  pravimo klika, če so si vsa vozlišča iz te množice med seboj sosednja.

Klika  $C$  na neusmerjenem grafu  $F$  je največja (maximal), če ne obstaja klika  $D$  na grafu  $G$ , tako da  $C \subseteq D$  in  $C \neq D$ . Klika  $C$  imenujemo maksimalna (maximum) klika, če ne obstaja

klika na grafu  $G$ , ki bi vsebovala več vozlišč kot največja klika  $C$ . Število vozlišč v maksimalni kliki imenujemo klično število (clique number) in ga označimo z  $\omega(G)$ .

Matrika sosednosti grafa  $G$  je kvadratna matrika  $A_G$ , kjer je  $(A_G)_{i,j} = 1$ , če  $(i, j) \in E$ , sicer pa  $(A_G)_{i,j} = 0$ .

Na matriko sosednosti v neusmerjenem grafu lahko gledamo kot na matriko podobnosti v problemu razvrščanja in posledično lahko uporabimo dominantno množico da najdemo združbe znotraj grafa.

Glede na povezavo z teorijo optimizacije, upoštevamo naslednji kvadratični program

$$\begin{aligned} \max f_\alpha(x) &= x^T (A_G + \alpha I) x \\ \text{p.p.} \quad x &\in \Delta \subset \mathbb{R}^n \end{aligned}$$

Kjer je  $I$  identična matrika,  $\alpha$  realno število in  $\Delta$  simpleks.

### Izrek2

naj bo graf  $G$  neusmerjen z matriko sosednosti  $A_G$  in naj bo  $0 < \alpha < 1$ . Vsaka največja klika  $C$  grafa  $G$  je dominantna množica od  $A_\alpha = A_G + \alpha I$ . Obratno, če je  $C$  dominantna množica od  $A_\alpha$  potem je  $C$  največja klika v  $G$ .