Razvrščanje z dominantnimi množicami - povzetek

Taja Debeljak, Anže Marinko
Finančni praktikum
Finančna matematika, Fakulteta za matematiko in fiziko

Jesen 2017

1 Uvod

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 $G = (V, E, \omega)$, kjer je $V = 1, \dots, n$ končna množica vozlišč, $E \subseteq V \times V$ množica usmerjenih povezav in $\omega : E \to \mathbb{R}$ 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 $A_{i,j} = \omega(i,j)$ 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

Definicija 1.1 *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*

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

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

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

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 elementa i glede na množico C definiramo kot

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

Vrednost $W_C(i)$ 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 1.2 *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 $i \notin C$*

2 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

$$max f(x) = x^T Ax \quad (1)$$
$$\text{p. p. } x \in \Delta \subset \mathbb{R}^n \quad (2)$$

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 $\sigma(x) = \{i \in V : x_i > 0\}$.

Definicija 2.1 *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

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

Za dominantno množico lahko torej vedno definiramo otežen vektor.

Izrek 1 *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 $i \notin \sigma$.*