# Razvrščanje z dominantnimi množicami

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 $V = 1, \ldots, n$ 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

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

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*

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

*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 $i \notin \sigma$.*

# 3 Povezava s teorijo grafov

Naj bo $G = (V, E)$ neusmerjen graf, kjer je $V = 1, 2, \ldots, n$ množica vozlišč in $E \subseteq V \times V$ množica povezav v grafu. Dve vozlišči $u, v \in V$ sta sosednji, 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 $C \neq D$. Kliko 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 neusmrejnem 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

$$max f_\alpha(x) = x^T(A_G + \alpha I)x \tag{3}$$

$$\text{p. p. } x \in \Delta \subset \mathbb{R}^n \tag{4}$$

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

**Izrek 2** *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.*

# 4 Razvrščanje z uporabo dmoinantnih množic

Naivna strategijabi lahko bilo oštevilčenje vseh podmnožic $C \subseteq V$ preverjanje pogojev iz Definicije 1. Ta rešitev je očitno precej neučinkovita, zato si bomo ogledali dve alternativni strategiji. Obe rešitvi izvirata iz teorije iger.

## 4.1 Dinamika replikatorjev

Dinamika replikatorjev (RD) je deterministična dinamična igra, ki je bila razvita v evolucijski teoriji iger. Ta teorija izvira iz zgodnjih sedemdesetih let kot poskus uporabe načel in orodja teorije iger v biološke namene za model evolucije živali. Predvideva idealni scenarij, s katerim so posamezniki ponavljajoč naključno prirejeni iz velike, idealno neskončne, populacije v igro dveh igralcev. V nasprotju s klasično teorijo iger se tukaj igralci naj ne bi obnašali racionalno ali naj ne bi imeli popolnih informacij o igri, ampak delujejo v skladu s podedovanim vedenjskim vzorcem ali čisto strategijo, in domneva se, da zaradi evolucijskega izbora proces sčasoma deluje na porazdelitev vedenja. Splošni razred evolucijskih enačb je podan z naslednjim nizom navadnih diferencialnih enačb:

$$\dot{x}_i = x_i g_i(x) \tag{5}$$

za $i = 1, \ldots, n$, kjer pika pomeni odvod po času in je $g = (g_1, \ldots, g_n)$ funkcija z odprtim definicijskim območjek, ki vsebujeΔ. Tukaj funkcija $g_i$ določa stopnjo, po kateri čista strategija i replicira. Običajno je potrebno, da je funkcija g strogo naraščajoča, kar pomeni, da je Lipschitz kontinuiran in da je $g(x)^T x = 0$

za vsak $x \in \Delta$. Zgornji pogoj zagotavlja, da ima sistem diferencialnih enačb (5) enolično določeno rešitev za vsako začetno populacijsko stanje. Slednji pogoj, namesto tega zagotavlja, da je množica $\Delta$ zaprta za sistem diferencialnih enačb (5), saj katerikoli evolucijski vzorec, ki se začne v $\Delta$ ostane v $\Delta$.
Za točko x velja, da je stacionarna (ali ravnovesna) točka za naše dinamične sisteme, če je $\dot{x}_i = 0$ (i $\in$ S). Stacionarna točka x je stabilna, če za vsako okolico U od x obstaja okolica V od x, tako da $x(0) \in V$ implicira $x(t) \in U$ za vse $t \geq 0$. Stacionarna točka naj bi bila asimptotično stabilna, če obstaja pot, ki se začne v njeni okolici in k njej konvergira, ko $t \to \infty$.
Dinamika monopolne igre izplačil predstavlja širok razpon običajne izbirne dinamike, za katero držijo koristne lastnosti. Intuitivno, za monopolno dinamiko izplačil se bodo strategije povezane z višjimi izplačili povečale pri višji stopnji. Formalno naj bi bila navadna dinamika izbora (5) monopolno-izplačilna, če $g_i(x) > g_j(x) \Leftrightarrow (Ax)_i > (Ax)_j$ za vse $x \in \Delta$ in $i, j \in V$.
Čeprav ta razred vsebuje veliko različnih dinamik, se izkaže, da si delijo veliko skupnih lastnosti. Za začetek, vsi imajo enako množico stacionarnih točk, saj je $x \in \Delta$ stacionarna točka pod kakršno koli izplačilno-monopolno dinamiko natanko tedaj, ko velja $(Ax)_i = x^T Ax$ za vsak $i \in \sigma(x)$.
Dobro znan podrazred izplačilno-monopolnih dinamičnih iger je podan z

$$\dot{x}_i = x_i[\phi((Ax)_i) - \sum_{j \in V} x_j \phi((Ax)_j)]$$

kjer je $\phi(u)$ naraščajoča funkcija u. Ti modeli se pojavijo v modeliranju evolucije vedenja z imitacijskimi procesi, kjer igralci občasno dobijo priložnost za spremembo svoje strategije.
Ko je $\phi$ identična funkcija, to je $\phi(u) = u$ , dobimo standardne enačbe za replikator v odvisnosti od časa,

$$\dot{x}_i = x_i[(Ax)_i - x^T Ax] \tag{6}$$

katerih osnovna ideja je, da je povprečna stopnja povečanja $\dot{x}_i/x_i$ enaka razliki med povprečno sposobnostjo strategije i in povprečno sposobnostjo celotne populacije.
Drug znan model se pojavi, ko je $\phi(u) = e_k u$, kjer je k pozitivna konstanta. Ko je k blizu 0, ta dinamika prestopi v standarden model replikatorja prvega reda upočasnjenega za faktorj k; poleg tega, pa se za velike vrednosti k model približuje tako imenovani dinamiki najboljšega odgovora.
Dinamika replikatorjev in bolj v splošnem kakršno koli izplačilo monopolne dinamike imajo nasledne lastnosti:

**Izrek 3** *V kakršni koli izplačilno-monopolni dinamiki velja:*

- *Nashevo ravnovesje je stacionarna točka;*
- *strogo Nashovo ravnovesje je asimptotično stabilno;*
- *stacionarna točka $x^*$, ki je meja notranje orbite, t.j. da je $\sigma(x(t)) = V$ za vsak $t \geq 0$ in $lim_{t \to \infty} x(t) = x^*$, je Nashevo ravnovesje;*
- *stabilna stacionarna točka je Nashevo ravnovesje;*
- *evolucijsko stabilna strategija (ESS) je asimptotično stabilna.*

V splošnem trditve iz izreka 4 ne veljajo v obratni smeri.
Poleg tega, če se omejimo na simetrične matrike izplačil, t.j., $A = A^T$, potem veljajo močnejše lastnosti:

**Izrek 4** *Če je $A = A^T$, potem velja naslednje:*

- *$x^T Ax$ je strogo naraščajoča vzdolž katere koli nekonstantne poti po (6). Z drugimi besedami, za vse $t \geq 0$ imamo $\frac{d}{dt}[x(t)^T Ax(t)] > 0$, razen če je $x$ stacionarna točka. Poleg tega vsaka taka pot konvergira proti (enolično določeni) stacionarni točki;*
- *$x$ je asimptotično stabilna natanko tedaj, ko je $x$ ESS.*

Da bi lahko izvajali dinamiko replikatorja v času, lahko uporabimo neko iterativno metodo, npr. metodo Runge-Kutta, za iskanje približka rešitev navadnih diferencialnih enačb. Druga možnost je, da lahko sprejmemo diskretno časovno podobo (6), ki je znana kot dinamika diskretnega replikatorja po času, ki jo dobimo z $x_i(t+1) = x_i(t)\frac{(Ax(t))_i}{x(t)^T Ax(t)}$, za $i \in V$. Opazimo, da je dinamika diskretnega replikatorja po času enostavno-invariantna, če je A nenegativna. To pa ni omejitev, ker vsaka izplačila A ohranijo svoje ravnotežje s konstantnim premikom. Torej, če ima A negativne elemente, lahko skonstruiramo matriko $\bar{A} = A + \mu E$ s pozitivnimi elementi z ustrezno izbiro $\mu > 0$, kjer je E matrika s samimi enicami.
Ker so ESS asimptotično stabilne v dinamiki replikatorjev po izrekih 3 in 4, lahko uporabimo to dinamiko za pridobivanje dominantnih množic, ki so v resnici v skladu z ESS, kot je bilo predhodno opisano. Poleg tega, če prevzamemo simetrične funkcije izplačil, spet opazimo povezavo s teorijo optimizacije, iz izhaja iz izreka 4, da dinamika replikatorjev lokalno maksimira $x^T Ax$. Končno, motivirana s povezavo s teorijo grafov, je bila dinamika replikatorjev uporabljena tudi za problem največje teže klika.

# 5 Dinamika okužb in imunizacije

Dinamika okužb in imunizacije je razred dinamike diskretnega časa, ki je bila uvedena, da bi premagali nekatere računalniške probleme s standardno evolucijsko dinamiko, ki vključujejo dinamiko replikatorjev. Naj omenimo nekaj problemov evolucijske dinamike kot dinamike replikatorjev: ima kvadratno kompleksnost prostora / časa na iteracijo, niso vse stacionarne točke Nasheva ravnovesja, konvergirajo le v limiti po številu iteracij in odkrivanje podpore ravnovesja je nerodno, če se dinamika ustavi pred pravilno konvergenco (tj. po končnem številu korakov).
Dinamika okužb in imunizacije je naslednja:

$$x(t+1) = \delta_{S(x)}(x)[S(x) - x] + x \tag{7}$$

, kjer smo zapisali x za x (t). Funkcija $S : \Delta \to \Delta$ je funkcija izbire strategije, ki izpolnjuje naslednje lastnosti:

$$S(x) = \begin{cases} y & \text{za nek } y \in \Upsilon(x), \text{ če je } \Upsilon(x) \neq \emptyset, \\ x & \text{sicer,} \end{cases} \tag{8}$$

kjer $\Upsilon(x)$ predstavlja množico tako imenovanih infektivnih strategij za x in je definirana kot $\Upsilon(x) = \{y \in \Delta : (y-x)^T Ax > 0\}$.
Dokler obstaja $y \in \Upsilon(x)$, strategija x ne more biti ravnovesna po definiciji infektivne strategije. Če je temu tako, se dinamika (7) zmeša x in y (okuži x z y), dokler ne pride do kršitve Nashevih pogojev, ki ga povzroči y, kar se zgodi z upoštevanjem mešalnega faktorja $\delta y(x)$, ki je podan z

$$\delta y(x) = \begin{cases} min\{\frac{(x-y)^T Ax}{(y-x)^T A(y-x)}, 1\}, & \text{če} (y-x)^T A(y-x) < 0, \\ 1, & \text{sicer.} \end{cases} \tag{9}$$

Dinamika okužb in imunizacije prinaša fiksno točko, ko je $\Upsilon(x) = \emptyset$. Če je temu tako, je tudi x Nashevo ravnovesje:

**Izrek 5** *Naj bo $x \in \Delta$ strategija. Potem so naslednje izjave ekvivalentne:*

1. *če $\Upsilon(x) = \emptyset$, za x ni infektivne strategije;*
2. *x je Nasheva strategija;*
3. *x je fiksna točka za dinamiko (7).*

**Dokaz:** Omejimo se na dinamiko replikatorja. Če se omejimo na simetrične matrike izplačil, imamo povprečno izplačilo strogo naraščajoče vzdolž katere koli nekonstantne poti dinamike okužb in imunizacije:

**Izrek 6** *Naj bo $\{x(t)\}_t \geq 0$ pot po (7). Torej za vse $t \geq 0$, $u(x(t+1), x(t+1)) \geq u(x(t), x(t))$ z enakostjo natanko tedaj, ko $x(t) = x(t+1)$, pod pogojem, da je matrica izplačila simetrična.*

**Dokaz:** Odvisno od izbire S(x) v (7) lahko dobimo drugačno dinamiko. V preostalem delu bomo pregledali še eno posebno, ki je preprosta in vodi do lepih lastnosti. Naj bo $u \in argmax_{j \in V}(Ax)_j$ in $v \in argmin_{j \in \sigma(x)}(Ax)_j$ in naj bo

$$z^+ = e^u \text{ in } z^- = \frac{x_v}{1 - x_v}(x - e^v) + x,$$

kjer je $e^u$ ničelni vektor z u-tim elementom 1. Funkcija izbire čiste strategije izbirna $S_{\text{čista}}$ je definirana kot

$$S_{\text{čista}}(x) = \begin{cases} z^+ & \text{če} (z^+ - z^-)^T Ax > 0 \\ z^- & \text{če} (z^+ - z^-)^T Ax < 0 \\ x & \text{sicer.} \end{cases}$$