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 množice v gruče (cluster) tako, da so si predmeti znotraj iste gruče bolj podobni med sabo, kot so si podobni z elementi iz ostalih gruč.
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 povečati 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 $\mathbb{R}^n$.

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

Izrek 1 *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_α potem je C največja klika v G.*

4 Povezava s teorijo iger

Teorija iger (za razliko od teorije optimizacije) lahko obravnava tudi nesimetrične matrike. Ideja je v zasnovi simetrične igre »clustering game« med dvema igralcema. Akcije V so čiste strategije, ki so na voljo igralcem, matrika A pa predstavlja njihova izplačila. Oba igralca imata popolno znanje o poteku igre in sprejemata neodvisne odločitve o tem katero strategijo bosta izbrala. Matrika izplačil predstavlja prihodke, ki jih dobi igralec če izbere določeno čisto strategijo. Torej, če igralca 1 in 2 uporabita strategijo $(\delta(i), \delta(j))$, kjer je $(i, j) \in V \times V$

potem igralec 1 dobi izplačilo $A_{i,j}$, igralec 2 pa $A_{j,i}$. Mešana strategija $x \in \Delta$ je verjetnost porazdelitve čistih strategij, ki modelira stohastično strategijo igre nekega igralca. Če igralec 1 in 2 igrata mešani strategiji $(x_1, x_2) \in \Delta \times \Delta$ potem sta pričakovani izplačili igralcev $x_1^T A x_2$ in $x_2^T A x_1$. Nashevo ravnovesje je profil mešane strategije $(x_1, x_2) \in V \times V$ pri katerem se noben od igralcev ne more povečati svojega izplačila ob nespremenjeni igri drugih igralcev, torej

$$y_1^T A x_2 \leq x_1^T A x_2 \text{ in } y_2^T A x_1 \leq x_2^T A x_1$$

Za vse $(y_1, y_2) \in \Delta \times \Delta$. Nashevo ravnovesje je simetrično, če velja $x_1 = x_2$. V primeru simetričnega Nashevega ravnovesja se prejšni neenakosti združita v

$$y^T A x \leq x^T A x \text{ za vse } y \in \Delta$$

. Z vidika *clustering game* je simetrično ravnovesje tisto, kjer imata oba igralca enake hipoteze o članstvu v gruči in noben od igralcev ne želi iz gruče. Še več, prejšni pogoj pomeni:

$$\begin{cases} (Ax)_i = x^T A x; & i \in \sigma(x) \\ (Ax)_i \leq x^T A x; & i \notin \sigma(x) \end{cases}$$

ki pa ga lahko interpretiramo kot notranjo homogenost gruče predstavljeno s podporo $\sigma(x)$ od x-a, za pričakovano podobnost katerega koli elementa iz gruče z nekim drugim elementom bo enaka.

Vendar pa Nash ne nujno zagotavlja pogoj maksimuma, kar lahko vseeno zagotovimo z izboljšanjem Nashevega ravnovesja poznanega kot *evolucijsko stabilna strategija (ESS)*. Simetrično Nashevo ravnovesje $x \in \Delta$ je evolucijsko stabilna strategija, če poleg pogoja $y^T A x \leq x^T A x$ za vse $y \in \Delta$ zadostuje tudi pogoju $y^T A x = x^T A x \Rightarrow y^T A x < x^T A x$ za vse $y \in \Delta \setminus x$.

Izrek 2 *Naj bo A matrika podobnosti pri primeru problema združevanja in naj bo Γ clustering game. Če je C dominantna množica v A, potem je vektor x^C ESS od Γ. Obratno, če je vektor x ESS od Γ, potem je $C = \sigma(x)$ dominantna množica od A, če $(Ax)_i \neq x^T A x$ za vse $i \notin C$.*

5 Razvrščanje z uporabo dominantnih množic

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

5.1 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:

$$\dot{x}_i = x_i g_i(x) \tag{3}$$

za $i = 1, \ldots, n$, kjer je x funkcija odvisna od t (čas).

5.2 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:

$$x(t+1) = \delta_{S(x)}(x)[S(x) - x] + x \tag{4}$$

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

$$S(x) = \begin{cases} y & \text{za nek } y \in \Upsilon(x), \text{ če je } \Upsilon(x) \neq \emptyset, \\ x & \text{sicer,} \end{cases} \tag{5}$$

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

$$\delta y(x) = \begin{cases} min\{\frac{(x-y)^T Ax}{(y-x)^T A(y-x)}, 1\}, & \text{če} (y-x)^T A(y-x) < 0, \\ 1, & \text{sicer.} \end{cases} \tag{6}$$

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

Izrek 3 *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 4 *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

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

Ker je iskanje prostora za infektivno strategijo zmanjšano z Δ do končne množice, ni očitno, da je $S_{\text{čista}}(x)$ dobro definirana izbirna funkcija, to je, da izpolnjuje (8). Vendar pa se lahko dokaže, da obstaja infektivna strategija za x natanko tedaj, ko je $S_{\text{čista}}(x)$ infektivna za x.
Druga lastnost, ki velja za čisto dinamiko okužb in imunizacije, ki si jo deli tudi z dinamiko replikatorjev, je karakterizacija ESS ravnovesja v smislu asimptotično stabilnih točk dinamike pri simetričnih matrikah izplačil.

Izrek 5 *Stanje x je asimptotično stabilno za dinamiko okužb in imunizacije s funkcijo$S_{čista}$ kot funkcijo izbire strategije natanko tedaj, ko je x ESS, pod pogojem, da je matrika izplačil simetrična.*

Dokaz. Ta izbirna funkcija kaže lepo lastnost, da je večanje kompleksnosti na iteracijo linearno tako v prostoru kot v času, v nasprotju z dinamiko replikatorja, ki ima kvadratno kompleksnost prostora / časa na iteracijo.

5.3 Iskanje več gruč

Dinamika replikatorja in dinamika okužb in imunizacije omogočata, da po konvergenci najdemo eno prevladujočo množico. Načeloma je cilj označiti vse dominantne množice za primere težav s problematiko združevanja, vendar je to v splošnem računsko intenzivno, ker imamo lahko eksponentno veliko gruč. V praksi je nedokončana, vendar dobra, pokritost dominantnih množic navadno zadostna za namene uporabe. Preglejmo nekatere računske strategije, ki so pomembne za dinamiko okužb in imunizacije, za (delno) označevanje dominantnih množic.

5.3.1 Strategija z več začetki

Prva strategija je zelo naiva in je sestavljena iz ponovnega zagona dinamike iz različnih, naključnih točk v simpleksu. Ta strategija je učinkovita, če so podatki sestavljeni iz nekaj gruč z velikimi bazeni privlačnosti, tako da je verjetnost vzorčenja točke, ki pripada posameznemu bazenu, relativno visoka. Jasno je, da ni nobenega zagotovila, da bo ta postopek večkrat izpisal iste gruče. Zato

je korak po obdelavi namenjen prepoznavanju in izločevanju ponavljajočih se gruč. Strategija z več začetki je dolgoročno optimalna, ker bo na koncu naštela vse dominantne množice, vendar pa lahko zahteva zelo veliko število vzorcev, kar je v praksi neučinkovito.

5.3.2 Strategija luščenja

Ta strategija sledi drugačni filozofiji in se uporablja predvsem v primerih, ko nas ne zanima iskanje prekrivajočih se gruč. Ideja je spet zelo preprosta in je sestavljena iz iterativnega izvlečenja dominantne množice in odstranitve njenih točk. S tem je očitno zagotovljeno, da se nobena dominantna množica ne bo večkrat pridobivala. Vendar ta strategija ni optimalna, ker je na eni strani število dominantnih množic, ki jih je potencialno mogoče označiti, navzgor omejeno s skupnim številom točk; po drugi strani pa, razen prve dominantne množice, vsi naslednje niso nujno dominantne množice prvotnega problema razvrščanja, saj se vsakič, ko so točke dominantne množice odvzete, spremeni osnovni problem združevanja. Kljub pomanjkanju optimnosti je strategija luščenja računalniško privlačna in se jo v praksi uporabljala v številnih aplikacijah.

5.3.3 Tabu-seznam strategije

Zadnja strategija temelji na dinamiki okužb in imunizacije. Spet bomo preprosto skicirali splošno idejo. Strategija tabu-seznamov sledi cilju nepopolnega nabiranja dominantnih množic, ki so podobne strategiji luščenja, vendar kljub temu zagotavlja dobro pokritost in je brez žrtvovanja optimalne rešitve, kar pomeni, da so izločene gruče dominantne množice prvotnega problema z gručami. Zamisel je ohraniti tabu-seznam med procesom razvrščanja, to je seznamom točk, ki jih ne bi smeli upoštevati med izločanjem dominantne množice. V nekaterih pogledih to počne tudi strategija luščenja, saj so vse točke že izločenih dominantnih množic dodane na tabu-seznam. Vendar, da bi preprečili možnost pridobivanja neveljavnih dominantnih množic, strategija tabu-seznamov preveri veljavnost domnevnih dominantnih množic, ki jih najdemo pod omejitvijo tabu-seznamov. V vsaki dominantni množici izločitvene iteracije je dominantna množica $C \subset V$ ugotovljena z izvajanjem dinamike okužb in imunizacij z naključno inicializacijo z omejitvijo, da ni mogoče uporabiti nobene točke v tabu-seznamu $T \subset V$, to je $C \subseteq V \setminus T$. Ta omejitev preprečuje, da bi bila C veljavna dominantna množica na splošno, to pomeni, da C ne bi bila dominantna množica glede na celotno množico točk V. Da bi preverili veljavnost domnevne dominantne množice C, se dinamika okužb in imunizacij ponovi, inicializira karakteristični vektorj x^C in dovoljuje uporabo vseh točk (tudi tistih v tabu-seznamu). Naj $C' \subseteq V$ naj bo dominantna množica, ki je veljavna, vendar morda ni nova. Obstajata torej dva možna scenarija, v katerih C' lahko štejemo za novo veljavno dominantno množico, ki se doda končnemu rezultatu razvrščanja:

1. $C' \nsubseteq T$,
2. $C' \subseteq T$, toda C' prej še nismo našli. V prvem primeru se tabu-seznam posodablja z dodajanjem katere koli točke iz $C' \setminus T$ (npr. naključne), medtem ko ostane v prvem primeru tabu-seznam tak, kot je. Če se namesto tega izkaže, da je $C' \subseteq T$, toda je C' že bilo najdeno, zavrnemo C' in dodamo v tabu-seznam katero koli izmed točk iz $V \setminus T$.

Strategija tabu-seznam zagotavlja, da se bodisi veljavna in nova dominantna množica pridobi pri vsaki iteraciji, ali pa se seznam tabu poveča za eno, s čimer se omeji število iteracij na $max(|V|, \#\text{najdenih gruč})$. Upoštevajmo, da bi lahko na časovno tabelo dodali še več točk, da bi pospešili proces združevanja, pri čemer bi sčasoma našli manjše število dominantnih množic.