

RAZVRŠČANJE EGO-CENTRIČNIH OMREŽIJ

Simona Korenjak-Černe in Vladimir Batagelj

Univerza v Ljubljani, EF, Kardeljeva ploščad 17, Ljubljana, Slovenija

FMF, Oddelek za matematiko, Jadranska 19, Ljubljana, Slovenija

Simona.Cerne@uni-lj.si

Vladimir.Batagelj@uni-lj.si

Povzetek

V sociologiji se pogosto srečamo z *ego-centričnimi* omrežji, kjer so enote analize izbrane osebe (*egi*) s svojimi osebnimi omrežji. Izbrane osebe so opisane z izbranimi spremenljivkami, ki so navadno merjene v različnih merskih lestvicah. Ostale osebe iz omrežja so prav tako opisane s (pogosto drugimi) spremenljivkami, ki opisujejo bodisi njihove lastnosti ali pa njihovo razmerje – povezavo z egom [7]. Taka omrežja lahko sestavlja tudi več tisoč enot.

V prispevku predstavljamo možnost zmanjšanja velikosti omrežij z razvrščanjem. Za opis ego-centričnih omrežij uporabimo simbolne objekte [1, 2], kjer vrednosti posamezne spremenljivke v enoti ali skupini enot povzamemo z njihovo porazdelitvijo. Tako dobimo natančnejši opis enot, kot pri dosedanjih pristopih, ki temeljijo na predstavitvi s srednjimi vrednostmi. Za razvrščanje simbolnih objektov smo razvili prilagojeno *metodo voditeljev* [6, 3, 4], za manjša omrežja pa tudi *hierarhično metodo združevanja*.

Rezultate omenjenih metod bomo predstavili na omrežju socialne opore, ki je bilo zbrano na Fakulteti za družbene vede v Ljubljani leta 2000 [5].

Summary

CLUSTERING EGO-CENTERED NETWORKS – *Ego-centered networks are frequently encountered in the social science research. A unit of the analysis is a respondent (ego) with his/her personal network (alters). Every unit is described by selected variables, which are usually measured in different scales. Several (other) variables are also measured on alters, that describe relationship among ego and alters and properties of alters [7]. Personal networks can be very large.*

In the paper we are presenting an approach how to reduce the size of such a network with clustering methods. For the descriptions of personal networks symbolic objects [1, 2] are used. Using variable's distributions (instead of the value of an appropriate statistics – e.g. mean value used in the 'standard' approach), the symbolic objects provide a more detailed description of personal networks. Based on this descriptions the adapted version of the leaders method [6, 3, 4] was developed, and for smaller networks also agglomerative hierarchical method.

The results of the proposed approach will be presented on the social support networks in Ljubljana 2000 dataset, collected at the Faculty of social sciences of University of Ljubljana [5].

1. EGO-CENTRIČNA OMREŽJA

V sociologiji se pogosto srečamo z *ego-centričnimi* omrežji, kjer so enote analize izbrane osebe (*egi*) s svojimi osebnimi omrežji. Izbrane osebe so opisane z več spremenljivkami, ki opisujejo lastnosti enot ali pa njihov odnos do ostalih oseb v njihovem osebnem omrežju, s katerimi so te izbrane enote povezane. Spremenljivke so navadno merjene v različnih merskih lestvicah (imenski, urejenostni, intervalni, razmernostni). Tudi ostale osebe iz omrežja so prav tako opisane s (pogosto drugimi) spremenljivkami, ki opisujejo bodisi njihove lastnosti ali pa njihovo razmerje – povezavo z egom [7]. Taka omrežja so lahko zelo velika, saj lahko obsegajo več sto egov in več tisoč oseb, s katerimi so povezani. Velikost omrežij onemogoča njihovo preglednost, zato v prispevku predstavljamo možnost zmanjšanja velikosti omrežij z razvrščanjem.

2. PREDSTAVITEV OMREŽIJ S SIMBOLNIMI OBJEKTI

Najpogosteje se za predstavitev ego-centričnih omrežij uporabljajo srednje vrednosti spremenljivk (težišče, mediana ipd.), pri tem pa je izgubljen precejšen del informacij o omrežju. Zato v prilagojenih metodah razvrščanja uporabimo raje bolj natančen opis, ki temelji na porazdelitvah vrednosti spremenljivk.

Naj bo \mathbf{E} končna množica izbranih oseb – egov in \mathbf{A} množica z njimi povezanih oseb. Vsakemu egu $\mathbf{X} \in \mathbf{E}$ pripada skupina z njim povezanih oseb $\mathbf{A}(\mathbf{X})$. Naj bo vsak ego $\mathbf{X} \in \mathbf{E}$ opisan z m_e spremenljivkami $U_1, U_2, U_3, \dots, U_{m_e}$

$$\mathbf{X} = [u_1, u_2, u_3, \dots, u_{m_e}],$$

kjer je $u_i = U_i(\mathbf{X})$ vrednost i -te spremenljivke U_i ega \mathbf{X} . Nadalje naj bo vsaka oseba \mathbf{Y} , povezana z egom, opisana z m_a spremenljivkami $V_1, V_2, V_3, \dots, V_{m_a}$

$$\mathbf{Y} = [v_1, v_2, v_3, \dots, v_{m_a}],$$

kjer je $v_i = V_i(\mathbf{Y})$.

Skupino oseb, povezanih z egom \mathbf{X} , predstavimo s simbolnim objektom $So(\mathbf{A}(\mathbf{X}))$, ki ga za vsako spremenljivko V_j sestavljajo porazdelitve p_j , ki so določene z vrednostmi, ki jih za spremenljivko V_j zavzamejo z egom povezane enote.

$$So(\mathbf{A}(\mathbf{X})) = [\mathbf{p}] = [p_1, p_2, p_3, \dots, p_{m_a}]$$

$$p_j \equiv p(\mathbf{X}; V_j) = [p(1, \mathbf{X}; V_j), p(2, \mathbf{X}; V_j), p(3, \mathbf{X}; V_j), \dots, p(k(V_j), \mathbf{X}; V_j)]$$

Zalogo vsake spremenljivke V torej razdelimo v izbrano število razredov $k(V)$. Število p_i v opisu z egom \mathbf{X} povezanih oseb $\mathbf{A}(\mathbf{X})$ predstavlja relativno frekvenco

$$p(i, \mathbf{X}; V) = \frac{q(i, \mathbf{X}; V)}{\text{card } \mathbf{A}(\mathbf{X})}, \quad i = 1, \dots, k(V),$$

kjer je $q(i, \mathbf{X}; V) = \text{card } Q(i, \mathbf{X}; V)$ in

$$Q(i, \mathbf{X}; V) = \{ \mathbf{Y} : \mathbf{Y} \text{ je oseba, povezana z egom } \mathbf{X}, \\ \text{katere vrednost pri spremenljivki } V \text{ leži v razredu } i \}.$$

Končno združimo opisa ega in skupine z njim povezanih oseb v razširjen *simbolni objekt* [1, 2]

$$So(\mathbf{X}) = [\mathbf{X}, \mathbf{p}],$$

ki je osnova simbolne analize podatkov ego-centričnih omrežij.

Oglejmo si predstavitev še na primeru:

Bazo podatkov, na kateri smo uporabili predstavljeni metodi, predstavlja omrežje socialne opore, ki je bilo zbrano na Fakulteti za družbene vede v Ljubljani leta 2000 [5]. Med 38 izbranimi spremenljivkami egov so tudi naslednje tri:

$U_1 =$ <i>zadovoljen z materialno oporo</i>	$U_2 =$ <i>zakonski stan</i>
1 = zelo nezadovoljen	1 = samski
2 = precej nezadovoljen	2 = živi kot poročen
3 = malo nezadovoljen	3 = poročen
4 = malo zadovoljen	4 = ločen
5 = precej zadovoljen	5 = ovdovel
6 = zelo zadovoljen	

$U_3 =$ <i>koliko let živi v mestu</i>	
1 = 10 ali manj	6 = 31-35
2 = 11-18	7 = 36-40
3 = 19-21	8 = 41-47
4 = 22-25	9 = 48-54
5 = 26-30	10 = 55 ali več

Spremenljivka U_1 je urejenostna, U_2 imenska in U_3 razmernostna.

Med desetimi spremenljivkami, izbranimi za osebe, povezane z egom, smo za primer izbrali naslednji dve:

$V_2 =$ <i>Spol</i>	$V_3 =$ <i>Kako daleč od ega živi</i>
1 = moški	1 = 15 min ali manj
2 = ženski	2 = nad 15 do 30 min
	3 = nad 30 do 60 min
	4 = uro ali dlje
	5 = živita skupaj

Opis enega izmed 1033 egov, ki ga v tem primeru označimo z $X1001$, naj bo

$$\begin{aligned} X1001 &= [4, 1, 4] = \\ &= [[0, 0, 0, 1, 0, 0], [1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]], \end{aligned}$$

kar pomeni, da je izbrana oseba – ego $X1001$ 'precej zadovoljna' z materialno pomočjo, ki ji jo nudijo z njo povezane osebe, da je 'samska' in da živi v mestu že 'od 22 do 25 let'.

Za tri osebe $Y1001:01$, $Y1001:02$ in $Y1001:03$, povezane z izbranim egom $X1001$, imamo naslednje podatke

$$\begin{aligned} Y1001:01 &= [2, 2] \\ &= [[0, 1], [0, 1, 0, 0, 0]] \\ Y1001:02 &= [2, 3] \\ &= [[0, 1], [0, 0, 1, 0, 0]] \\ Y1001:03 &= [1, 2] \\ &= [[1, 0], [0, 1, 0, 0, 0]] \end{aligned}$$

Torej je skupina, sestavljena iz teh treh oseb, predstavljena s simbolnim objektom kot

$$So(A(X1001)) = \left[\left[\frac{1}{3}, \frac{2}{3} \right], \left[0, \frac{2}{3}, \frac{1}{3}, 0, 0 \right] \right]$$

Oba opisa združimo v simbolni objekt ega $X1001$

$$So(X1001) = \left[4, 1, 4, \left[\frac{1}{3}, \frac{2}{3} \right], \left[0, \frac{2}{3}, \frac{1}{3}, 0, 0 \right] \right]$$

3. PRILAGOJENI METODI RAZVRŠČANJA

Za opisano predstavitev skupin definiramo različnost med egoma \mathbf{X}_1 and \mathbf{X}_2 kot uteženo vsoto različnosti glede na posamezne spremenljivke

$$d(\mathbf{X}_1, \mathbf{X}_2) = \sum_{j=1}^m \alpha_j d(\mathbf{X}_1, \mathbf{X}_2; V_j), \quad \sum_{j=1}^m \alpha_j = 1,$$

kjer je

$$d_1(\mathbf{X}_1, \mathbf{X}_2; V) = \frac{1}{2} \sum_{i=1}^{k(V)} |p(i, \mathbf{X}_1; V) - p(i, \mathbf{X}_2; V)| \quad (1)$$

ali

$$d_2(\mathbf{X}_1, \mathbf{X}_2; V) = \frac{1}{2} \sum_{i=1}^{k(V)} (p(i, \mathbf{X}_1; V) - p(i, \mathbf{X}_2; V))^2. \quad (2)$$

Z $\alpha_j \geq 0$ ($j = 1, \dots, m$) označimo uteži, ki so lahko za vse spremenljivke enake, lahko pa so tudi različne, če imamo o pomembnosti spremenljivk še kakšne dodatne informacije.

Na osnovi tako izbranih različnosti definiramo problem razvrščanja kot optimizacijski problem, pri katerem iščemo porazdelitev \mathbf{C}^* , pri kateri je

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C}).$$

Za kriterijsko funkcijo smo uporabili njeno najpogostejšo obliko

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, L_C),$$

kjer je L_C voditelj (predstavnik) skupine C . Voditelji so v našem primeru tako kot enote predstavljeni s simbolni objekti – porazdelitvami.

Metoda voditeljev, ki smo jo prilagodili za razvrščanje prej opisanih simbolnih objektov [6], predstavlja izpeljanko t.i. dinamične metode razvrščanja [3], katere posebna izpeljanka je tudi v statistiki zelo pogosto rabljena k-središčna metoda.

Na kratko lahko opišemo metodo voditeljev z naslednjo shemo:

```

L := L0; C := C0;
repeat
    L := G(C);
    C := F(L);
until voditelji se ustalijo
    
```

kar pomeni, da na začetku izberemo začetne voditelje in začetno razbitje, nato pa ponavljamo izbiro novih voditeljev (preslikava G) in nato vsako enoto pridružimo njej najbližjemu novemu voditelju (preslikava F). Ponavljanje končamo, ko se voditelji ne spreminjajo več.

Izkaže se, da so optimalni voditelji za kriterijsko funkcijo z različnostjo d_1 (1) določeni z *maksimalnimi frekvencami*, za drugo različnost d_2 (2) pa s *povprečjem relativnih frekvenc*. Tako določeni voditelji nam omogočajo preproste razlage karakteristik dobljenih skupin.

Poleg metode voditeljev smo za razvrščanje predstavljenih simbolnih objektov razvili tudi *hierarhično metodo združevanja*, kjer skupine v vozliščih hierarhije predstavimo s simbolnimi objekti – porazdelitvami. Tudi pri tej metodi uporabimo iste definicije različnosti med skupinami, kot smo jih uporabili pri metodi voditeljev za različnosti med egi. Prednost hierarhičnih metod pred nehierarhičnimi je pregledna drevesna predstavitev, osnovna pomanjkljivost te in vseh ostalih hierarhičnih metod pa je, da so primerne le za razvrščanje nekaj sto enot, ker uporabljajo matriko različnosti med enotami. Zato je tudi ta metoda primerna le za manjša omrežja. Tako kot pri metodi voditeljev imamo tudi pri tej metodi za vsako skupino shranjeno informacijo o celotnih porazdelitvah vrednosti enot oz. egov, ki so v njej.

3.1. Rezultati razvrščanja

Omrežje socialne opore, nad katerim smo uporabili opisani metodi, je bilo zbrano na Fakulteti za družbene vede v Ljubljani med marcem in junijem leta 2000 s telefonskimi intervjuji (CATI). Reprezentativen vzorec sestavlja 1033 prebivalcev Ljubljane – egov, opisanih z 38 spremenljivkami, merjenimi v različnih merskih lestvicah. Oseb, povezanih z egi, je bilo skupno 5849, opisane pa so bile z 10 dodatnimi spremenljivkami [5].

Izhodne datoteke programov razvrščanja vsebujejo veliko informacij o skupinah, ki jih dobimo pri razvrščanju, npr. število enot – egov v skupini, 'napako' skupine, ki je izračunana kot vsota različnosti med enotami in voditeljem skupine, voditelju najbližji in najbolj različen ego z ustreznima različnostma. Po želji lahko zahtevamo izpis frekvenc za vsako spremenljivko, vrednosti značilnih razredov, t.j. razredov z največjo frekvenco in izpis vseh enot v skupini.

Primer informacij o eni od skupin:

SKUPINA 11: število enot = 45

spremenljivka = DQ3A(socialna opora)

max frekvenca = 35 (77.78 %), manjkajočih: 0, nesmiselnih: 0

q(C;V): 0 0 0 0 10 35

maxL(V): 0 0 0 0 0 1

Razlaga: V skupini 11 je 45 egov. Pri spremenljivki DQ3A(socialna opora) ni bilo manjkajočih ali nesmiselnih vrednosti. Deset egov je 'precej zadovoljnih' s tovrstno oporo, večina egov v tej skupini (natančneje 35, kar znaša 77.78 %) pa je 'zelo zadovoljna' s socialno oporo, ki jim jo nudijo z njimi povezane osebe.

Nekatere značilnosti skupine:

91.11 % D2(št. otrok v gospodinjstvu, mlajših od 19 let) = 0

57.78 % D4(koliko let živi v mestu) = od 22 do 25

86.67 % D7(starost) = od 22 do 26

86.67 % D8(poklic) = študent

91.11 % D9(izobrazba) = 4-letna srednja šola

95.56 % D10(zakonski stan) = samski
75.56 % SPOL(spol) = ženski
88.89 % DQ1A(materialna opora) = zelo zadovoljen
80.00 % DQ2A(informacijska opora) = zelo zadovoljen
77.78 % DQ3A(socialna opora) = zelo zadovoljen
84.44 % DQ4A(čustvena opora) = zelo zadovoljen
55.37 % Q12(spol osebe, povezane z egom) = ženski
52.44 % Q14(kako daleč od ega živi) = največ 15 min

Iz zgornjega opisa vidimo, da so v skupini večinoma samske osebe (95.56 % egov v skupini je samskih) s 4-letno srednjo šolo (91.11 %), 75.56 % je žensk, 86.67 % je študentov, večinoma so zelo zadovoljni z oporo, ki jim jo nudijo z njimi povezane osebe (spremenljivke DQ1A, DQ2A, DQ3A, DQ4A), ki so v 55.37 % ženske in živijo zelo blizu njih samih.

4. ZAKLJUČEK

Razvrščanje egov iz ego-centričnih omrežij kot simbolnih objektov predstavlja nov pristop v tovrstnih analizah, saj v nasprotju z dosedaj uveljavljenimi predstavami omrežij s srednjimi vrednostmi spremenljivk omogoča ohranitev večjega dela informacije, vključitev vseh spremenljivk v samo analizo in tudi vnos dodatnega znanja preko vrednosti uteži. Razvrščanje je časovno sprejemljivo tudi za zelo velika omrežja, spremenljivke, ki opisujejo lastnosti enot ali njihove odnose do ostalih oseb pa so lahko merjene v poljubni merski lestvici.

Literatura

- [1] Bock H.H., Diday E. (eds.): *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer, Heidelberg, 2000.
- [2] Diday E.: *Extracting Information from Extensive Data sets by Symbolic Data Analysis.* Indo-French Workshop on Symbolic Data Analysis and its Applications, Paris, 23-24. September 1997, Paris IX, Dauphine, p. 3-12.
- [3] Diday E.: *Optimisation en classification automatique*, Tome 1.,2.. INRIA, Rocquencourt, 1979, (in French).
- [4] Hartigan J.A.: *Clustering Algorithms.* Wiley, New York, 1975.
- [5] Kogovšek T., Ferligoj A., Coenders G.: *Estimating reliability and validity of personal support measurements: full information ML estimation with missing data planned by design.* Paper presented at SMABS 2000.
- [6] Korenjak-Černe S., Batagelj V.: *Clustering large datasets of mixed units*, in: *Advances in Data Science and Classification.* Rizzi, A., Vichi, M., Bock, H.-H. (Eds.), Springer, 1998.
- [7] Müller C., Wellman B., Marin A.: *How to use SPSS to study ego-centered networks.* To appear in *Bulletin de méthodologies sociologiques*, BMS, No. 64, 1999.