Uporabna matematika, četrti kolokvij 2007/08.

3. 6. 2008

Pri analizi skupine povezanih grafov smo pridobili naslednje podatke:

Ozn.	$\lvert V\rvert$	$\lvert E\rvert$	$s = \Delta - \delta$	$\zeta = \lvert E\rvert - \lvert V\rvert + 1$	Skupina
P_3	4	3	1	0	Pot
C_4	4	4	0	1	Cikel
G_1	4	5	2	2	Kaktus
T_1	5	4	2	0	Drevo
G_2	5	5	2	1	Kaktus
G_3	5	6	2	2	Kaktus
P_5	6	5	1	0	Pot
C_6	6	6	0	1	Cikel
T_2	6	5	2	0	Drevo
G_4	8	9	3	2	Kaktus
T_3	10	9	3	0	Drevo
C_{10}	10	10	0	1	Cikel
G_5	10	11	1	2	Kaktus
G_6	11	12	8	2	Kaktus
C_{11}	11	11	0	1	Cikel
G_7	11	12	3	2	Kaktus

Naloga 1 (25) *Z uporabo podatkov iz gornje razpredelnice odgovorite na naslednja vprašanja:*

(a) Prikažite graf, ki ustrezata podatkovni hierarhiji atributa s. [3]

(b) Prikažite graf, ki ustrezata podatkovni hierarhiji atributa ζ. [3]

(c) Prikažite graf, ki ustrezata podatkovni mreži atributov s, ζ. [6]

(d) K vsakemu vozlišču grafa iz (c) pripišite ustrezno podatkovno kocko. [15]

Rešitev.

- (a) Ker atribut s ni sestavljen, je graf njegove hierarhije sestavljen iz dveh vozlišč $\emptyset$ in s ter povezave med njima. Pri tem oznaka $\emptyset$ ustreza ekvivalenčni relaciji na domeni atributa s, katere edini ekvivalenčni razred je celotna domena, s pa ekvivalenčni relaciji na domeni atributa s, katere ekvivalenčni razredi so vsi singeltoni.
- (b) Ker atribut ζ ni sestavljen, je graf njegove hierarhije sestavljen iz dveh vozlišč $\emptyset$ in ζ ter povezave med njima. Pomen oznak $\emptyset$ in ζ je podoben, kot v (a).

- (c) Graf je kartezični produkt grafov iz (a) in (b), torej cikel C_4, njegova vozlišča pa so (povezana v tem vrstnem redu) $(\emptyset, \emptyset)$, $(\emptyset, \zeta)$, (s, ζ), $(s, \emptyset)$.
- (d) Vozliščem pripadajo naslednje podatkovne kocke:

(s,ζ)	0	1	2
0	0	4	0
1	2	0	1
2	2	1	2
3	1	0	2
8	0	0	1

$(\emptyset,\zeta)$	0	1	2
	5	5	6

$(s,\emptyset)$	
0	4
1	3
2	5
3	3
8	1

$(\emptyset,\emptyset)$	
	16

□

Naloga 2 (25+10) *[15] Kakšna je absolutna in kakšna je relativna napaka pri napovedovanju atributa* `Skupina` *z enomestnim pravilom*

- *z atributom ζ in*
- *z atributom s?*

[10] Kateri nabor (večmestnih) klasifikacijskih pravil napove vrednost atributa skupina brez napake? [BONUS 10] Katero od klasifikacijskih pravil iz prejšnjega nabora je izrek v teoriji grafov? Izberite enega in ga dokažite.

Rešitev. V razpredelnici je podatkovna kocka za atributa ζ in `Skupina` kombinirana z absolutno napako e za napovedovanje atributa `Skupina` z enomestnim pravilom na osnovi atributa ζ:

	0	1	2
Pot	2	0	0
Drevo	3	0	0
Cikel	0	4	0
Kaktus	0	1	6
e	2	1	0

Skupno napoved zgreši v treh primerih, absolutna napaka je torej 3 in relativna $\frac{3}{16}$.
V razpredelnici je podatkovna kocka za atributa s in `Skupina` kombinirana z absolutno napako e za napovedovanje atributa `Skupina` z enomestnim pravilom na osnovi atributa s:

	Pot	Drevo	Cikel	Kaktus	e
0	0	0	4	0	0
1	2	0	0	1	1
2	0	2	0	3	2
3	0	1	0	2	1
8	0	0	0	1	0

Skupno napoved zgreši v štirih primerih, absolutna napaka je torej 4 in relativna $\frac{1}{4}$. Boljši atribut za napovedovanje je torej ζ, ki nam da enomestna pravila za atribut `Skupina`:

- Če je $\zeta(G) = 0$, potem je graf drevo.
- Če je $\zeta(G) = 1$, potem je graf cikel.
- Če je $\zeta(G) = 2$, potem je graf kaktus.

Prvo pravilo ima dve izjemi, ko bi moralo povedati, da je graf pot. Pri obeh izjemah velja $s = 1$, medtem ko pri drugih primerih velja $s > 1$. Pravilo lahko torej popravimo s pomočjo atributa s. Podobno ima drugo pravilo eno izjemo, pri kateri velja $s > 0$, med tem ko je pri vseh ciklih $s = 0$. Tako dobimo naslednji nabor pravil, ki brez napake napovedo vrednost atributa s na danem primeru:

- Če je $\zeta(G) = 0$ in $s(G) = 1$, potem je graf pot.
- Če je $\zeta(G) = 0$ in $s(G) > 1$, potem je graf drevo.
- Če je $\zeta(G) = 1$ in $s(G) = 0$, potem je graf cikel.
- Če je $\zeta(G) = 1$ in $s(G) > 0$, potem je graf kaktus.
- Če je $\zeta(G) = 2$, potem je graf kaktus.

Ob predpostavki, da je G povezan, so prva tri pravila izreki v teoriji grafov. Dokažimo, da je tretje pravilo izrek. Za dani graf velja $\zeta(G) = |E| - |V| + 1 = 1$, torej graf ima enako mnogo vozlišč, kot povezav. Ker je $\Delta(G) - \delta(G) = 0$, je njegova maksimalna stopnja enaka minimalni stopnji. Graf je torej regularen; vsa vozlišča imajo enako stopnjo d. Vemo, da je vsota stopenj v grafu enaka dvakratniku števila povezav, torej v našem primeru $2|E| = d|V|$. Sledi $d = 2$. Dani povezani graf je torej regularen s stopnjo 2, torej je cikel. □

Naloga 3 (25) *Izračunajte entropijo atributa* `Skupina` *in informacijski doprinos atributa* ζ.

Rešitev. Podatkovna kocka atributa `Skupina` je

Pot	Drevo	Cikel	Kaktus
2	3	4	7

Entropija particije na $2, 3, 4, 7$ elementov je tako enaka

$$\text{ent}(2,3,4,7) = -\tfrac{2}{16}\log_2(\tfrac{2}{16}) - \tfrac{3}{16}\log_2(\tfrac{3}{16}) - \tfrac{4}{16}\log_2(\tfrac{4}{16}) - \tfrac{7}{16}\log_2(\tfrac{7}{16}) = 1,85.$$

S pomočjo podatkovne kocke za atributa `Skupina` in ζ iz prejšnje naloge izračunamo informacijsko vsebino atributa ζ:

$$\text{info}(\zeta, \text{Skupina}) = \tfrac{5}{16}\left(-\tfrac{2}{5}\log_2(\tfrac{2}{5}) - \tfrac{3}{5}\log_2(\tfrac{3}{5})\right) + \tfrac{5}{16}\left(-\tfrac{4}{5}\log_2(\tfrac{4}{5}) - \tfrac{1}{5}\log_2(\tfrac{1}{5})\right) + 0 = 0,53.$$

Informacijski doprinos atributa ζ je torej enak:

$$\text{doprinos}(\zeta, \text{Skupina}) = \text{ent}(\text{Skupina}) - \text{info}(\zeta, \text{Skupina}) = 1,85 - 0,53 = 1,32.$$

□

Naloga 4 (25) *V psevdokodi napišite algoritem za iskanje gruč s pomočjo k-median. Algoritem za* $k = 4$ *dvakrat izvedite na podatkih 1, 3, 4, 9, 11, 12, 24, 25, 27, 28, 34, 35, 39, 42: prvič za izhodišče vzemite podatke 11, 12, 25, 28, drugič pa podatke 24, 25, 27, 28. Komentirajte rezultata.*

Rešitev.

```
Naj bo P množica vseh podatkov.
Naj bo C=P.

za i=1 do k
  c[i]=naključen podatek iz C;
  C = C brez c[i];

Ponavljaj
  Za i od 1 do k
    v R[i] zberemo podatke, ki so najbližje točki c[i]
  Za i od 1 do k
    c[i] izračunamo kot mediano razreda R[i]
Dokler so bile v zadnjem koraku spremembe
```

Če začnemo s podatki $c_1 = 11$, $c_2 = 12$, $c_3 = 25$, $c_4 = 28$, potem je izhodiščna delitev enaka (1, 3, 4, 9, 11), (12), (24, 25), (27, 28, 34, 35, 39, 42). Mediane razredov po prvem koraku so enake $c_1 = 4$, $c_2 = 12$, $c_3 = 24,5$, $c_4 = 34,5$. Nova delitev na gruče je enaka (1, 3, 4), (9, 11, 12), (24, 25, 27, 28), (34, 35, 39, 42). Mediane so v naslednjem koraku enake $c_1 = 3$, $c_2 = 11$, $c_3 = 26$, $c_4 = 37$. Delitev na gruče je enaka prejšnji, torej smo zaključili postopek s končno delitvijo (1, 3, 4), (9, 11, 12), (24, 25, 27, 28), (34, 35, 39, 42).

Če začnemo s podatki $c_1 = 24$, $c_2 = 25$, $c_3 = 27$, $c_4 = 28$, potem je izhodiščna delitev enaka (1, 3, 4, 9, 11, 12, 24), (25), (27), (28, 34, 35, 39, 42). Mediane razredov po prvem koraku so enake $c_1 = 9$, $c_2 = 25$, $c_3 = 27$, $c_4 = 35$. Nova delitev na gruče je enaka (1, 3, 4, 9, 11, 12), (24, 25), (27, 28), (34, 35, 39, 42). Mediane so v naslednjem koraku enake $c_1 = 6,5$, $c_2 = 24,5$, $c_3 = 27,5$, $c_4 = 37$. Delitev na gruče je enaka prejšnji, torej smo zaključili postopek s končno delitvijo (1, 3, 4, 9, 11, 12), (24, 25), (27, 28), (34, 35, 39, 42).

Sklepamo, da je rešitev, ki jo predstavi postopek iskanja gruč s pomočjo k-median, odvisna od začetnih podatkov. Prva delitev se zdi boljša, torej je smiselno na začetku izbrati podatke, ki so ustrezno razpršeni. □