Istraživanje podataka 1 - zadatak

Zadatak: Data je matrica sličnosti skupa podataka. Izvršiti hijerarhijsko klasterovanje korišćenjem min veze. Rezulatat prikazati dendogramom.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Rešenje

Na početku svaka instanca predstavlja poseban klaster. U svakom koraku se spajaju dva najbliža, tj. najsličnija klastera. Pošto se primenjuje min veza, sličnost između dva klastera se određuje na osnovu dve najsličnije (najbliže) instance u različitim klasterima. Postupak se nastavlja dok sve instance ne budu u jednom klasteru. Spajanje izvršeno u svakom koraku je označeno crvernom bojom u pridruženom dendogramu.

- I korak - spajaju se klasteri sa instancama p2 i p5, pošto je ovaj par instanci najsličniji. Pre sledećeg spajanja, potrebno je izračunati sličnost (s) između novog klastera {p2,p5} i ostalih klastera primenom min veze. Kako se kao mera bliskosti koristi sličnost, najbliži par iz dva klastera (min veza, videti sliku **??**) će imati najveću sličnost.

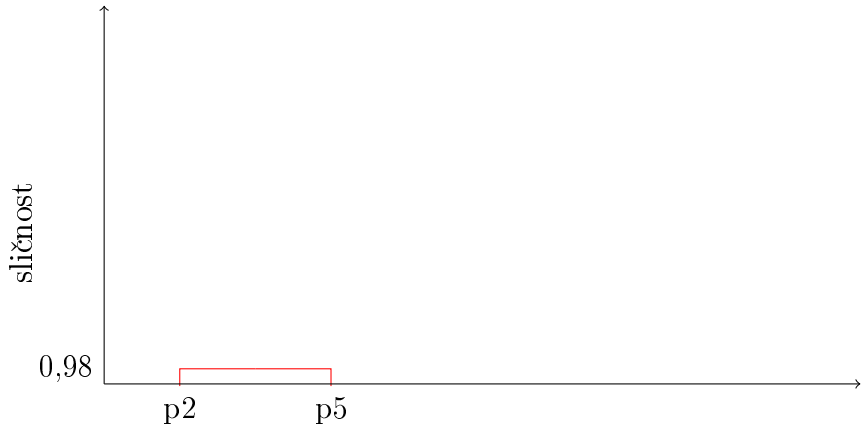
$$s(\{p2, p5\}, p1) = max(s(p2, p1), s(p5, p1)) = max(0, 1, 0, 35) = 0, 35$$
$$s(\{p2, p5\}, p3) = max(s(p2, p3), s(p5, p3)) = max(0, 64, 0, 85) = 0, 85$$
$$s(\{p2, p5\}, p4) = max(s(p2, p4), s(p5, p4)) = max(0, 47, 0, 76) = 0, 76$$

  Napomena: da je data matrica različitosti (umesto matrice sličnosti) i da se koristi min veza za određivanje bliskosti dva klastera, za najbliži par iz dva klastera bi važilo da imaju *najmanju* udaljenost.

  Nakon spajanja, matrica sličnosti klastera izgleda:

  a dendogram:

	p1	{p2,p5}	p3	p4
p1	1	0,35	0,41	0,55
{p2,p5}	0,35	1	0,85	0,76
p3	0,41	0,85	1	0,44
p4	0,55	0,76	0,44	1



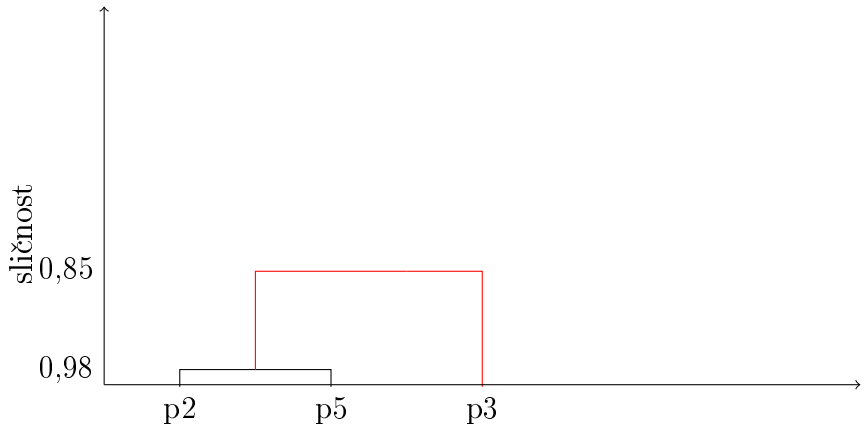
- II korak - najsličniji su klasteri {p3} i {p2, p5}, te se oni spajaju. Sličnost novog klastera sa ostalim klasterima je:

$$s(\{p2, p3, p5\}, p1) = max(s(\{p2, p5\}, p1), s(p3, p1)) = max(0,35, 0,41) = 0,41$$
$$s(\{p2, p3, p5\}, p4) = max(s(\{p2, p5\}, p4), s(p3, p4)) = max(0,76, 0,44) = 0,76$$

Nakon spajanja, matrica sličnosti klastera izgleda:

	p1	{p2,p3,p5}	p4
p1	1	0,41	0,55
{p2,p3, p5}	0,41	1	0,76
p4	0,55	0,76	1

a dendogram:



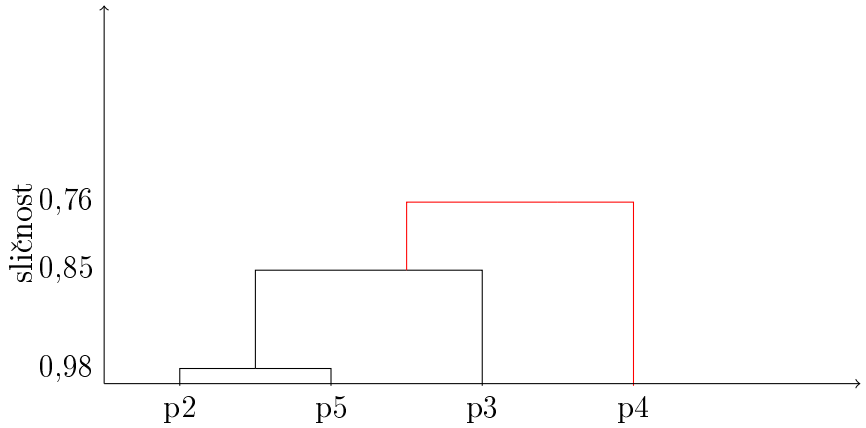
- III korak - najsličniji su klasteri {p4} i {p2, p3, p5}, te se spajaju u jedan. Sličnost poslednja dva klastera {p2, p3, p4 ,p5} i {p1} je

$$s(\{p2, p3, p4, p5\}, p1) = max(s(\{p2, p3, p5\}, p1), s(p4, p1)) = max(0,41, 0,55) = 0,55$$

Nakon spajanja, matrica sličnosti klastera izgleda:

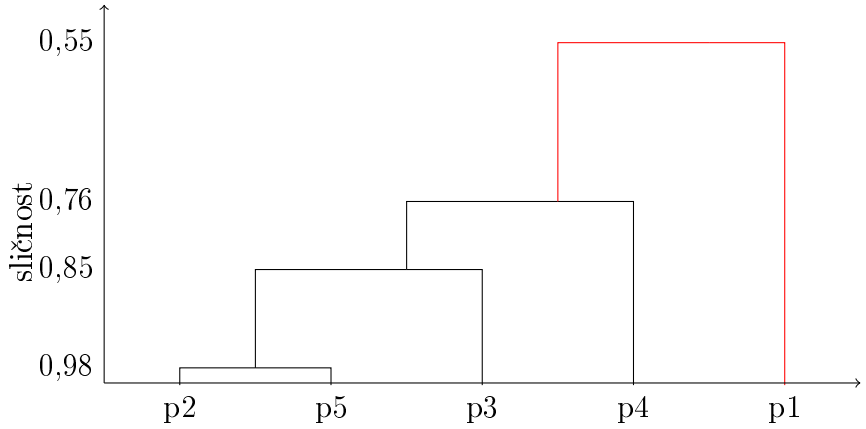
	p1	{p2,p3,p4,p5}
p1	1	0,55
{p2,p3, p4, p5}	0,55	1

a dendogram:



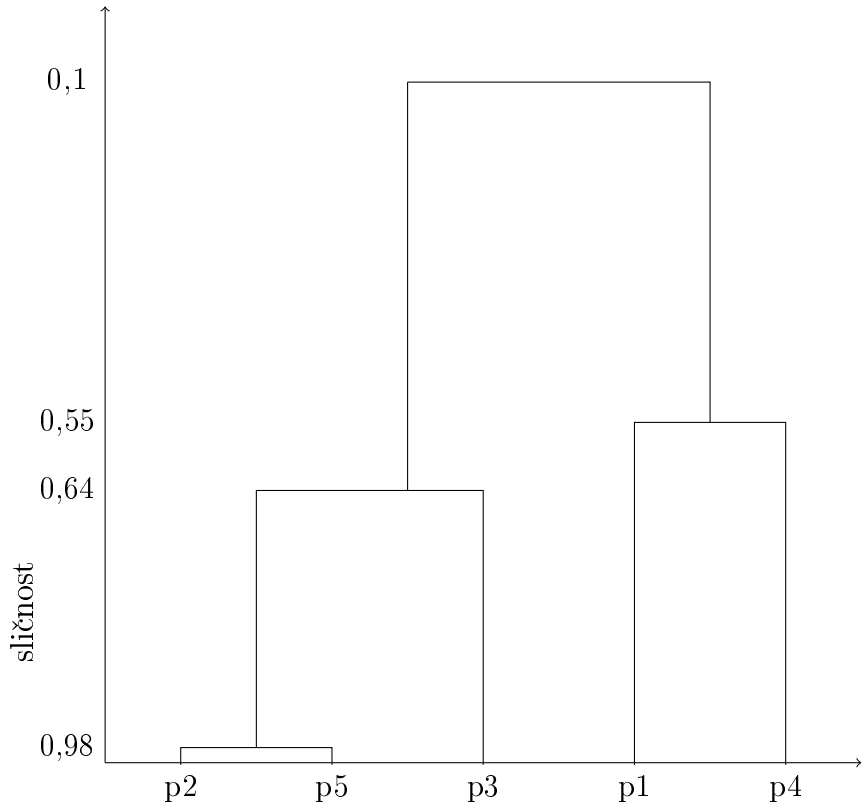
- IV korak - spajaju se poslednja dva klastera: klaster sa instancom p1 i klaster sa instancama {p2, p3, p4, p5}.

  Dendogram:

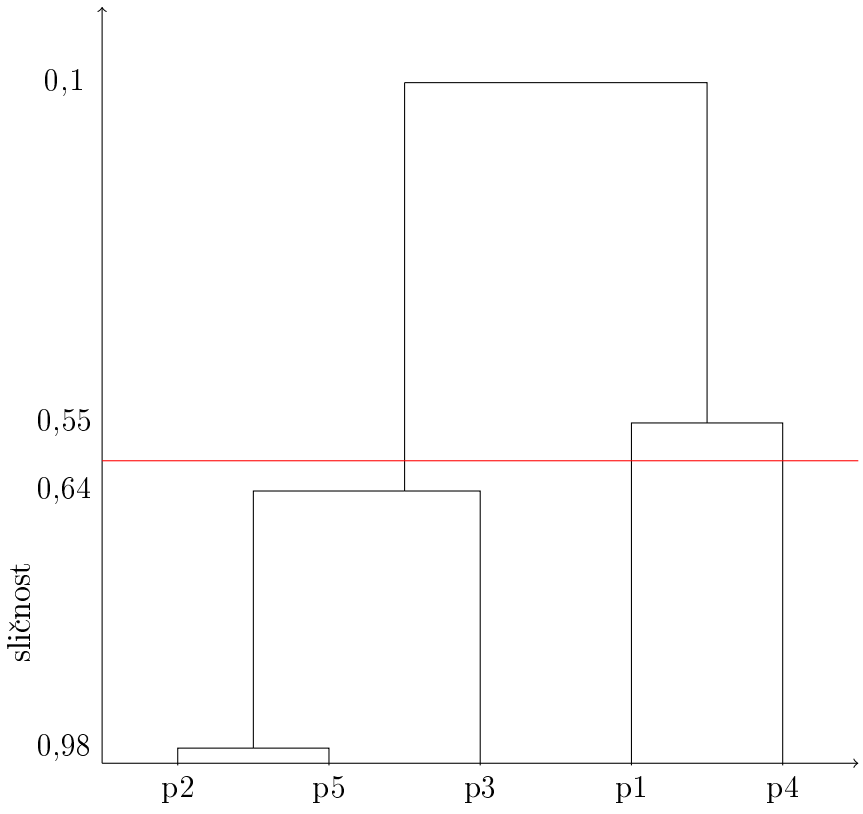


Ukoliko je potrebno izdvojiti dva klastera, poništava se poslednje spajanje i izdvajaju se klasteri {p1} i {p2, p3, p4, p5}. Ukoliko je potrebno izdvojiti tri klastera, poništavaju se poslednja dva spajanja i izdvajaju se klasteri {p1}, {p4} i {p2, p3, p5}.

Ako se umesto min veze koristi max veze pri računanju sličnosti dva klastera, traži se par instanci tih klastera sa najmanjom sličnošću i njihova sličnost je sličnost tih klastera. Dendogram klasterovanja sa max vezom izgleda:



Da je pre primene hijerarhijskog klasterovanja, kao kriterijum za zaustavljanje klasterovanja zadat prag za sličnost klastera 0,6, poslednja dva spajanja se ne bi izvršila i izdvojeni bi bili klasteri: {p4}, {p1} i {p2, p3, p5}. Na sledećem dendogramu je crvenom linijom prikazan zadati prag. Sva spajanja klastera čija je sličnost manja od praga se ne izvršavaju.


0,1
0,55
0,64
0,98
sličnost
p2
p5
p3
p1
p4