1 Uvod

Prva naloga pri predmetu Uvod v odkrivanje znanj iz podatkov je zadevala analizo glasovanja pri vsakoletnem glasbenem tekmovanju Evrovizija. Cilj naloge je bila aplikacija tehnike hierarhičnega gručenja nad podatki v priloženi CSV datoteki, ki je vsebovala podatke o glasovanju za tekmovanja med leti 1998 in 2009 ter vizualizacija rezultatov te metode z lastno implementacijo izrisa dendrograma.

1.1 Podatki

Podatki o glasovanju so bili podani v obliki CSV datoteke, ki je na voljo na portalu Kaggle. Datoteka vsebuje podatke za glasovanje med leti 1998 in 2008 in je organizirana tako, da so po vrsticah podane skladbe, ki so opisane z 63 atributi, ki med drugim zajemajo podatke o tem, kako so za skladbo v tej vrstici glasovale vse posamezne države, ki na tekmovanju lahko glasujejo.

Pri uporabi podatkov je bila potrebna pazljivost pri občasnih manjkajočih podatkih v stolpcih, ki vsebujejo podatke o glasovanju posameznih držav za posamezno skladbo. Teh manjkajočih vrednosti je bilo v celotni CSV datoteki 4612, kar predstavlja 33.7% vseh podatkov v teh stolpcih.

2 Metode

2.1 Podatki

Glavni vmesnik s problemsko domeno naloge je bil implementiran kot razred z imenom HierarchicalClustering v datoteki naloga.py. Ta razred enkapsulira metode, ki skupaj definirajo mehanizem za izvajanje hierarhičnega gručenja nad podatki, s katerimi je instanca tega razreda inicializirana. Predprocesiranje podatkov podanih v CSV datoteki je implementirano v funkciji *read_file*, ki se nahaja v isti datoteki kot omenjena definicija razreda. Ta funkcija podatke vsebovane v CSV datoteki pretvori v obliko vektorjev predstavljenih kot slovarji, ki preslikujejo imena držav, ki na tekmovanju lahko glasujejo, v tabelo numeričnih vrednosti, ki predstavljajo relativno frekvenco glasov, ki jih je ta država namenila ostalim nastopajočim državam.

Pri nalogi so bile torej analizirane podobnosti med državami v smislu podobnosti profila relativnega deleža glasov namenjenega ostalim državam oziroma regijam, ki jih države predstavljajo. Z metodo hierarhičnega gručenja so bile ustvarjene skupine držav z določeno stopnjo podobnosti profila glasovanj.

V isti datoteki je definirana še ena pomožna funkcija z imenom *get_labels*, ki generira matriko 2×47 z imeni držav ter njihovimi regijami. Ta matrika služi kot označevalka stolpcev v podatkovnih vektorjih, saj se v njej na istem stolpičnem indeksu nahaja ime države, katero numerična vrednost v podatkovnem vektorju zadeva, ter njena regija.

2.2 Računanje razdalj

Pri metodi hierarhičnega gručenja ločimo med merami razdalje med primeri ter med skupinami. v metodi *row_distance*, ki jo enkapsulira razred *HierarchicalClustering*, sta implementirani tako evklidska kot manhattanska razdalja na takšen način, da je možno enostavno spremeniti katera izmed njiju se uporablja. Pri reševanju naloge je bila za računanje razdalj med primeri (posameznimi vektorji) uporabljena evklidska razdalja.

Tudi v metodi za računanje razdalj med posameznimi skupinami (klustri) so implementirane vse pogosto uporabljene mere razdalj (*single linkage*, *complete linkage*, *average linkage* in *Ward distance*). Zopet je bila pozornost namenjena temu, da je možno enostavno spremeniti katera mera razdalje se pri reševanju naloge dejansko uporablja. Pri reševanju naloge je kot mera razdalje med skupinami bila uporabljena t.i. *average linkage*, ki predstavlja povprečno razdaljo med vsemi pari primerov iz teh dveh skupin. Pri vizualizaciji rezultatov je bila za namen primerjave uporabljena tudi Wardova razdalja, ki pogosto pri hierarhičnem gručenju deluje bolje kot povprečna razdalja.

Pri konstrukciji vektorjev ter računanju razdalj je bilo potrebna pozornost na v prejšnjem poglavju omenjene manjkajoče vrednosti v podatkih. Ker so kot numerične vrednosti v konstruiranih podatkovnih vektorjih bile uporabljene relativne frekvence glasov za posamezno državo se je na tem mestu kot naravna rešitev ponujalo upoštevanje le poznanih vrednosti pri konstrukciji vektorjev za opis glasovanja držav, saj je bil z uporabo normaliziranih vrednosti odstranjen vpliv absolutne frekvence glasov, ki bi se med posameznimi državami zaradi teh anomalij lahko razlikovala.

3 Rezultati

3.1 Izris dendrograma

Eden od ciljev naloge je bil grafična predstavitev rezultatov hierarhičnega gručenja v obliki za to nalogo razvite implementacije izrisa dendrograma. Dendrogram je v celoti sestavljen iz ASCII znakov. Implementiran je tako, da pri izrisu upošteva izračunane razdalje med primeri in skupinami. Dolžina črt, ki povezujejo dva primera oziroma dve skupini je torej proporcionalna razdalji med tema dvema primeroma oziroma skupinama.

Primer dendrograma, ki je nastal kot rezultat gručenja z uporabo povprečne razdalje kot mero razdaje med dvema skupinama, je prikazan na sledeči sliki.

Slika 1: Vizualizacija rezultatov gručenja z dendrogramom

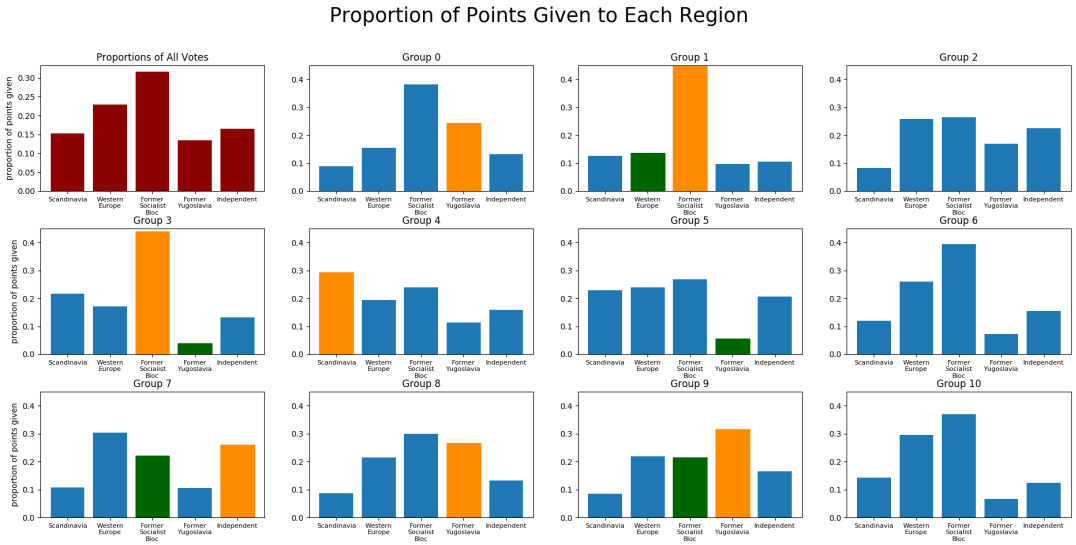
3.2 Skupine in njihove preferenčne izbire

Pri reševanju naloge so bile skupine sestavljene z metodo hierarhičnega gručenja. Vsaka država je bila predstavljena z vektorjem relativnih frekvenc glasov, ki jih je namenila nastopajočim državam. Metoda hierarhičnega gručenja je skupine konstruirala na podlagi podobnosti teh vektorjev.

Razporeditev držav v skupine je zapisana v spodnji tabeli. Preferenčne izbire skupin pa so prikazane na naslednji sliki. Za lažjo vizualizacijo so bile države grupirane v regije (pripadnost regiji je bil podatek vključen v CSV datoteko). V grafu je za vsako skupino predstavljena relativna frekvenca glasov, ki so jih države, ki sestavljajo posamezno skupino, namenile za države, ki pripadajo posamezni regiji. Vrednosti, ki od povprečja vseh skupin odstopajo za več kot 8% v katerokoli smer so posebej označene, in sicer z oranžno barvo za odstopanje navzgor ter z zeleno barvo za odstopanje navzdol. Pod sliko so navedene države, ki jih je hierarhično gručenje dodelilo v posamezno skupino ter njihove regije. Kot primerjava osnovnim rezultatom pridobljenim z uporabo povprečne razdalje med primeri kot mere razdalje med skupinami, ki so predstavljeni z dendrogramom, so tukaj prikazani tudi rezultati gručenja z uporabo Wardove razdalje, ki pogosto da boljše rezultate.

Skupina	Države v skupini
0	Monako, Slovaška, Azerbajdžan, San Marino, Češka, Gruzija, Srbija in Črna gora, Črna gora, Srbija
1	Belorusija, Ukrajina, Armenija, Moldavija, Madžarska
2	Albanija, Bolgarija, Romunija
3	Litva, Estonija, Latvija
4	Finska, Islandija, Danska, Norveška, Švedska
5	Malta, Irska, Združeno kraljestvo
6	Grčija, Ciper, Izrael
7	Francija, Nemčija, Belgija, Nizozemska
8	Hrvaška, Turčija
9	Bosna in Hercegovina, Slovenija, Švica, Avstrija, Makedonija
10	Andora, Portugalska, Španija, Poljska, Rusija

Tabela 1: Razporeditev držav v skupine (kot mera razdalje med skupinami je bila uporabljena Wardova razdalja)



Slika 2: Relativne frekvence glasov, ki so jih skupine namenile državam iz posamezne regije (kot mera razdalje med skupinami je bila uporabljena Wardova razdalja)

Skupina/Regija	Bivši socialistični blok	Zahodna evropa	Neodvisni	Bivša Jugoslavija	Skandinavija
povprečje	31.6%	23%	16.6%	13.5%	15.3%
0	38.2%	15.5%	13.2%	24.1% (+)	9%
1	53.6% (+)	13.6% (-)	10.5%	9.7%	12.6%
2	26.5%	25.7%	22.5%	17%	8.3%
3	44% (+)	17.1%	13.2%	4% (-)	21.7%
4	24%	19.3%	15.9%	11.4%	29.4% (+)
5	26.9%	23.9%	20.7%	5.5% (-)	23%
6	39.5%	26%	15.4%	7.1%	12%
7	22.2%	30.4%	25.9% (+)	10.6%	10.9%
8	30%	21.5%	13.2%	26.6% (+)	8.7%
9	21.4% (-)	22%	16.5%	31.5% (+)	8.5%
10	37%	29.6%	12.4%	6.7%	14.3%

Tabela 2: Tabelarični prikaz frekvenc glasov (Odstopanja od povprečja za več kot 8% so označena z znakom + oz. -)

4 Izjava o izdelavi domače naloge.

Domačo nalogo in pripadajoče programe sem izdelal sam.