3. domača naloga: razvrščanje v skupine

24. april 2018

1 Uvod

Podatki, ki jih dobimo na vhodu so podani s 61637 primeri, ki so opisani z 25133 atributi. Naša naloga je bila, da te podatke razvrstimo v skupine, pri čemer mora biti kvaliteta razvrstitve ocenjena z uporabi popravljeni indeks po Randu (ARI), vsaj 0,29. Največji izziv naloge je delati z velikimi podatki.

2 Metoda

Naloge sem se lotil tako, da sem najprej z *mmread()* prebral podatke iz datoteke, nato pa sem matriko prepisal tako, da sem ohranil samo prvih 2500 atributov, saj je tam matrika najbolj gosta. Podatke sem tudi binariziral, torej vse vrednosti, ki so bile večje od nič sem nastavil na ena. S tem sem dosegel, hitrejše delovanje K-means in boljše rezultate. Za razvrščanje v skupino sem uporabil metodo K-means, katerim sem nastavil število gruč na 45, kar se je izkazalo za najbolj optimalno. Inicializacijsko metodo sem pustil na privzeti vrednosti, *k-means++*, saj ta veliko hitreje konvergira, kot če bi začetne centre izbirali naključno. Rezultat, ki sem ga uspel dobiti s to metodo je ARI=0,21

3 Konsenzno razvrščanje

Konsenznega razvrščanja nisem implementiral.

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.