3. domača naloga: razvrščanje v skupine

24. april 2018

1 Uvod

Cilj naloge je bila podatke predstavljene v redki matriki razvrstiti v skupine. Naloga je bila sestavljena iz dveh delov. Prvi del je bila implementacija konsenznega razvrščanja. Drugi del je predstavljalo tekmovanje, kjer smo lahko uporabili algoritem po lastni izbiri.

2 Metoda

Algoritem, ki je dosegel najboljši rezultat je primere (torej vsako vrstico v redki matriki) normaliziral. Nato sem izbral 1000 stolpcev (atributov), ki imajo največ ne-ničelnih elementov. Na novo pridobljeno matriko sem ponovno normaliziral po primerih. Sledila je razvrščanje v skupine na podlagi algoritma KMeans++.

3 Konsenzno razvrščanje

Konsenzno razvrščanje je metoda, ki ocenjuje stabilnost skupin. Algoritem je dokaj splošen - določimo mu lahko način vzorčenja (*angl. resampling*) in algoritem za razvrščanje v skupine (*angl. clustering algorithem*). Algoritem na vzorcih opravi razvrščanje v skupine za različne število skupin K ($K \in 2, ..., K_{max}$). Za določen K se število razvrstitev v skupine ponavlja h-krat. Naj bo D_h vzorec pridobljen iz podatkov D. Na podlagi določenih skupin zgradimo matriko M_h, kjer je element v i-ti vrstici in v j-tem stolpcu enak 1, če i-ti in j-ti primer iz podatkov D pripadata isti skupini. Izračunamo tudi matriko I_h, ki je indikatorska matrika. Če sta i-ti in j-ti primer oba prisotna v vzorcu D_h je vrednost $I_h(i, j)$ enaka 1. Konsenzna matrika $C(i, j)$ je izračunana kot:

$$C(i,j) = \frac{\sum_h M_h(i,j)}{\sum_h I_h(i,j)}. \tag{1}$$

Sledi določitev optimalnega števila skupin $\hat{K}$. Za to potrebujemo empirično komutativno porazdelitev (CDF), ki je definirana kot:

$$CDF(c) = \frac{\sum_{i<j} 1\{C(i,j)\} \leq c}{N(N-1)/2}, \tag{2}$$

kjer je N število primerov v podatkih D. Za določeno konsenzno matriko C_K izračunamo območje pod CDF, ki je podan z naslednjo formulo:

$$A(K) = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(X_i), \tag{3}$$

kjer so $x_1, ..., x_m$ urejeni unikatni elementi konsenzne matrike C_K. Sedaj še določimo deležno spremembo območja pod CDF, ko se K povečuje, $\Delta(K)$, s formulo:

$$\Delta(K) = \begin{cases} A(K) & \text{if K} = 2 \\ \dfrac{A(K+1) - A(K)}{A(K)} & \text{if K} > 2 \end{cases} \tag{4}$$

Za $\hat{K}$ izberemo K kjer je $\Delta(K)$ največji. Primere iz D na koncu razdelimo v $\hat{K}$ skupin na osnovi $C_{\hat{K}}$.