

2. domača naloga: Informativnost značilk pri napovedovanju posameznih tem

Matic Potočnik (63060270)

5. marec 2012

1 Uvod

V nalogi smo predprocesirali in binarizirali podnabor podatkov iz tekmovanja JRS 2012 Data Mining Competition, ter ocenili kateri atributi nosijo največ informacije za posamezen razred, s pomočjo mere medsebojne informacije, ter permutacijskega testa.

2 Metode

2.1 Medsebojna informacija

Za ocenjevanje kakovosti atributov smo uporabili mero medsebojne informacije med atributom in razredom oz. informacijski prispevek[?].

Matematična definicija

Enačba za izračun informacijskega prispevka je:

$$\text{Gain}(A) = H_C - H_{C|A}$$

Pri tem je entropija H definirana kot:

$$\begin{aligned} H(p) &= p \cdot \log_2 \frac{1}{p} \\ H_{\mathbf{p}} &= \sum_i H(p_i) \\ H_{\mathbf{p}|\mathbf{q}} &= \sum_i q_i \sum_j H(p_j|q_i) \end{aligned}$$

Primer izračuna

Podano imamo tabelo z vrednostmi atributa in razreda:

atribut	razred
1	1
0	0
1	0
1	1
1	0

Iz tabele ugotovimo ustrezne verjetnosti in pogojne verjetnosti:

$$\begin{aligned}P(A = 0) &= \frac{1}{5} \\P(C = 0) &= \frac{2}{5} \\P(C = 0 \mid A = 0) &= 1 \\P(C = 0 \mid A = 1) &= \frac{1}{2}\end{aligned}$$

Sedaj lahko izračunamo informacijski prispevek:

$$\text{Gain}(A) = H\left(\frac{2}{5}, \frac{3}{5}\right) - \left(\frac{1}{5} \cdot H(1, 0) + \frac{4}{5} \cdot H\left(\frac{1}{2}, \frac{1}{2}\right)\right) = 0.17095059445466863$$

Zgornji rezultat smo izračunali po formuli s pomočjo programa Mathematica, zelo podoben rezultat pa vrne tudi funkcija `Orange.feature.scoring.InfoGain()`:

```
>>> Orange.feature.scoring.InfoGain(0, data)
0.17095059156417847
```

Razlika se pojavi na devetem mestu in jo gre najverjetneje pripisati temu, da Mathematica operira z višjo natančnostjo, kot to dopušča pythonski tip float, ki ga uporablja funkcija `Infogain`.

2.2 Permutacijski test

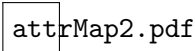
Permutacijski test uporabimo za bolj realno oceno kakovosti atributov pri napovedi. Metode za ocenjevanje kakovosti atributov nam dajo nek rezultat, ampak je ta relativen in včasih težko ugotovimo dejansko kvaliteto atributa in ali ga je vredno upoštevati pri nadaljni obravnavi.

Permutacijski test izvedemo tako, da z eno od metod za ocenjevanje kakovosti atributov ocenimo kakovost atributa za izbrani razred, nato pa podatke o razredu nekajkrat naključno permutiramo in izračunamo ocene kakovosti atributa nad permutiranimi podatki o razredu. Atribut je za razred informativen, če je njegova ocena kakovosti za ta razred, statistično značilno boljša od ocen, ki smo jih dobili pri naključnih permutacijah razreda.

3 Rezultati

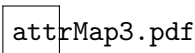
Uporabili smo permutacijski test z $n = 500$ ponovitvami in stopnjo značilnosti $\alpha = 0.05$. Izvajanje programa je trajalo 3 ure, 39 minut in 22 sekund na računalniku s procesorjem Core 2 Duo E8200 2.66GHz. Pri tem je bila uporabljena optimizacija, pri kateri je program preskočil atribut, če je naletel na več kot $n * \alpha = 25$ naključnih permutacij razreda z višjim informacijskim prispevkom kot ga je imel prvotni razred.

Iz prvega histograma lahko vidimo porazdelitev števila razredov glede na število informativnih atributov. Povprečen razred ima 608 informativnih atributov, sicer pa so vrednosti na intervalu od 215 do 1779 informativnih atributov na razred.

attrMap2.pdf

Slika 1: Porazdelitev števila informativnih atributov, glede na število razredov

Če statistiko obrnemo, opazimo da je vsak posamezen atribut informativen za več razredov. Tukaj razpon sega od 9 do 54 razredov na atribut, povprečje pa je pri 24.

attrMap3.pdf

Slika 2: Porazdelitev števila razredov, glede na število informativnih atributov

Največkrat (54-krat) so informativni atributi: 746, 919, 1203, 1414, 1636

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

Literatura

- [1] Kononenko Igor, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing, Chichester, UK, 2007.