Šesta domača naloga

Anže Pečar (63060257)

26. april 2012

1 Uvod

Namen šeste domače naloge je seznaniti se s podatki za tekmovanje iz področja kemoinformatike.

2 Podatki in opis problemske domene

Problemska domena ima 1776 atributov in 3751 primerov. Vrednosti atributov so med 0 in 1. Najbolj pogostih 5 vrednosti atributov je zapisanih v tabeli 1.

Tabela 1: Najbolj pogoste vrednosti atributov.

vrednost atributa	št ponovitev
0.0	5593148
1.0	421473
0.125	20583
0.1667	17677
0.25	17127

Večina atributov ima eno ali dve različni vrednosti. Kar precej (836) atributov ima samo dve različni vrednosti in bi jih lahko smatrali za diskretne atribute. Če bi za diskretne atribute vzeli atribute, ki imajo 5 različnih vrednosti, bi imeli 1095 diskretnih in 681 zveznih atributov.

3 Informativnost atributov

Za izračun informativnosti sem uporabil Orangeovo implementacijo Reliefa. Relief mi je za 3 najbolj informativne atribute izbral D27 (0.2145), D1036 (0.1368), D961 (0.1313).

Relief ocenjuje kvaliteto atributov glede na to, kako dobro ločujejo posamezne razrede. Za vsak primer Relief poišče dva najbližja soseda; enega iz istega razreda (najbližji zadetek) in enega iz različnega razreda (najbližja zgrešitev). Če imata dva primera različni vrednosti najbližjega zadetka atributa, potem ta atribut ločuje dve instanci z istim razredom, kar ni zaželjeno in se zato vrednost tega atributa zmanjša.

Algoritem relief računa razliko med dvema diskretnima atributoma po formuli:

$$diff(A, I_1, I_2) = \begin{cases} 0 & if value(A, I_1) == value(A, I_2) \\ 1 & drugace \end{cases}$$

Za zvezne atribute pa računa po formuli:

$$diff(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)}$$

kjer je A izbran atribut, I_1 in I_2 pa dva primera (instanci). Formuli sem povzel po članku [1]

Katere atribute upoštevati je močno odvisno od modela, ki ga uporabimo za učenje. Preizkusil sem več različnih pragovnih vrednosti z različnimi modeli. Pri naključnih drevesih sem dobil najboljši rezultat pri pragu 0.0, knn pa pri pragu 0.1.

4 Ocenjevanje kakovosti napovednih modelov

Za napovedovanje sem uporabil tri kvalifikacijske algoritme, ki so vključeni v Orangeovi knjižnici - RandomForest, kNN in NaivniBayes. Algoritme sem preizkusil z 10 kratnim prečnim preverjanjem, rezultati pa so podani v tabeli 2.

Tabela 2: Rezultati 10 kratnega prečnega preverjanja

algoritem	log loss ocena
RandomForest	0.464369549067
kNN	0.536824035629
NaivniBayes	0.627878923437

Na lestvici vodečih bi se z najboljšim rezultatom (RandomForest) uvrstili na 150. mesto. To je dober začetek, ampak za boljši rezultat bo potrebno implementirati stackanje algoritmov.

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

Literatura

[1] Ian H. Witten & Eibe Frank, *Theoretical and Empirical Analysis of ReliefF and RReliefF* Marko Robnik-Sikonja, Igor Kononenko, 2003.