Odločitvena drevesa

Bojan Leskošek

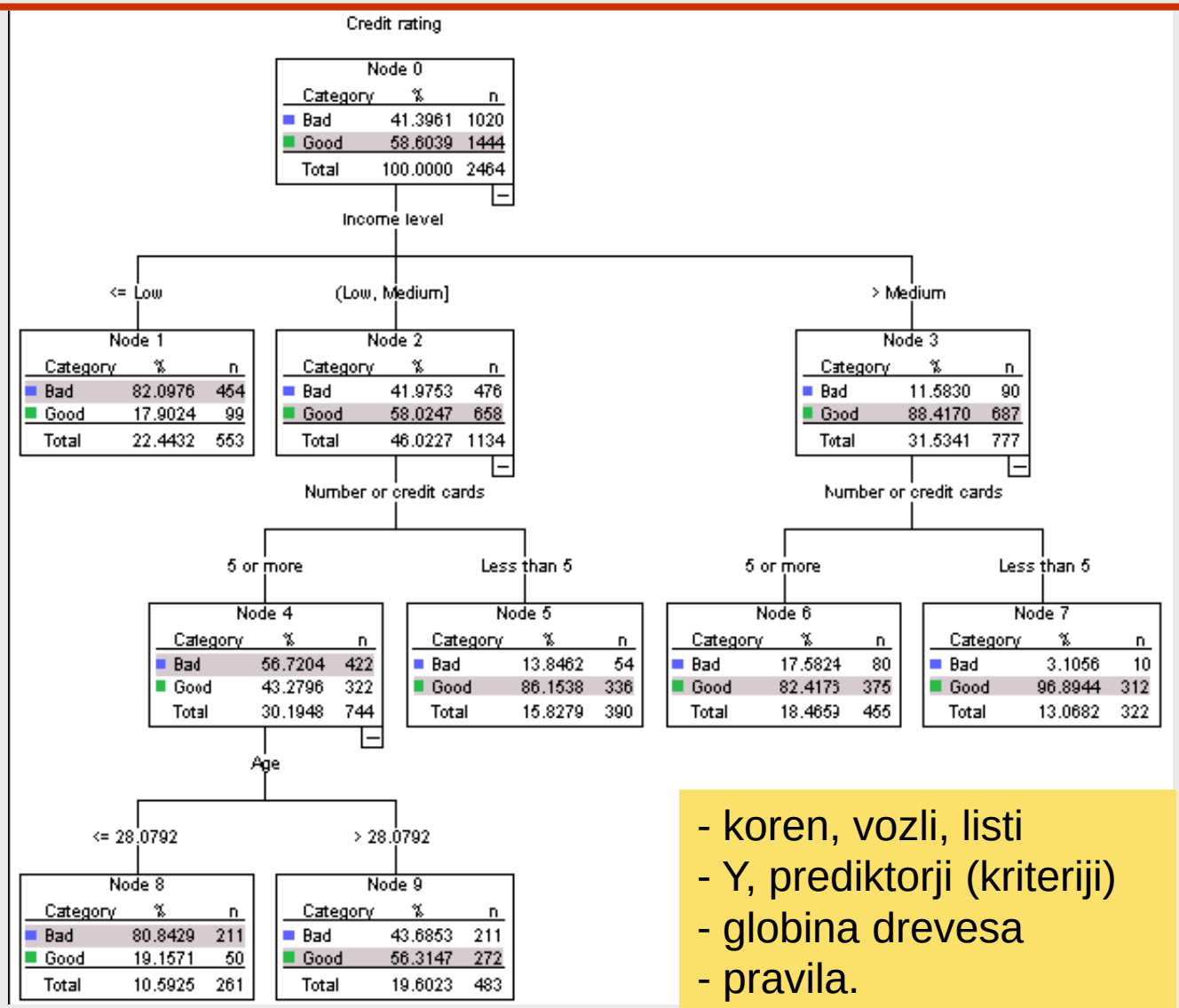
UL, Fakulteta za šport

IBMI, 16.3.2010

Kazalo

- Primeri
- Osnovni pojmi, vrste
- Namen, prednosti / slabosti
- Izgradnja OD:
  - metode rasti
  - vrednotenje, obrezovanje
- Softver
- Priporočeni viri.

Primer: posojilojemalci



- koren, vozli, listi
- Y, prediktorji (kriteriji)
- globina drevesa
- pravila.

Primeri

1. iskanje kandidatov z dobrimi možnostmi za uspeh v študiju / poklicu / športu
2. iskanje značilnosti, na osnovi katerih otrokom svetujemo izbiro športne panoge
3. iskanje lastnosti igralcev, ki so pomembne za razporejanje na igralna mesta
4. iskanje kriterijev, na osnovi katerih lahko ocenimo težavnost gimnastičnih prvin.

Kaj so odločitvena drevesa?

- Po namenu sorodna regresijskim modelom – običajni cilj je čim boljša napoved in razlaga odvisne spremenljivke Y
- Metoda ni definirana z modelom, ampak z algoritmom (*recursive partitioning algorithm*): dano množico enot se glede na vrednosti prediktorjev zaporedoma razvršča v vedno manjše, čim bolj čiste (po Y izenačene) skupine:
  - teži se, da je model čim bolj enostaven (drevo čim manjše) ob čim manjši izgubi informacije oz. tem bolj točni napovedi.
- Rezultat je v obliki drevesa ali niza pravil *če-potem*.

… kaj so odločitvena drevesa?

- Začetek v 60. letih pr. st. v statistiki (CHAID), od 70. intenziven razvoj v računalniški znanosti (avtomatsko učenje)
- Pravila se zgradijo "samodejno" na osnovi učnih primerov, preverjajo pa na osnovi novih primerov z neznano vrednostjo Y
  - Zaradi te samodejnosti in dobrega (včasih "preveč":) odkrivanja interakcij (2-, 3-, večsmernih) med prediktorji tudi alternativno ime "samodejni odkrivalec interakcij" (*AID- Automatic Interaction Detector*).

… kaj so odločitvena drevesa?

- OD (*decision trees*) je pravzaprav skupina metod, ki spadajo v dve kategoriji:
  - regresijska drevesa: Y je številska (razrezana v kategorije)
  - klasifikacijska drevesa: Y je (po naravi) kategorialna
- Metoda izgradnje je lahko enaka, razlikuje pa se kriterij ocene rezultata:
  - pri regresijskih d. običajno RSS
  - pri klasifikacijskih d. neka mera čistosti (*purity*) vozlov, npr. *Ginijev indeks* ($1-\Sigma p_i^2$), devianca ali entropija.

Namen OD

- napovedovanje
- redukcija števila prediktorjev
- ugotavljanje interakcij (iskanje podskupin s specifičnimi povezavami)
- klasifikacija, segmentacija
- diskretizacija Y, združevanje kategorij Y.

Prednosti / slabosti

- Prednosti:
  - enostavnost (pravila), nazornost (grafična predstavitev),
  - splošna uporaba (različni merski nivoji, tudi manjkajoči podatki, ne-aditivni modeli),
  - nezahtevne predpostavke,
  - učinkovitost na velikem podatkovju s kompleksno strukturo (interakcije, nelinearne povezave)
  - redukcija prediktorjev
    - nepomembni prediktorji izločeni iz drevesa
    - posamezno enoto ni treba izmeriti z vsemi prediktorji.

… prednosti / slabosti

- Slabosti:
  - *overfitting* – pravila se preveč prilagajajo (redkim) podatkom, zato obsežna drevesa, ki jih ni možno razložiti, povezati s teorijo. Zdravila:
    - skrben izbor prediktorjev
    - dovolj velik vzorec enot (učnih primerov)
    - *obrezovanje* (*prunning*)
    - preverjanje
  - pristranost – prediktorji z veliko kategorijami (zlasti nominalni) lahko dobijo prevelik pomen.

Podatki v OD

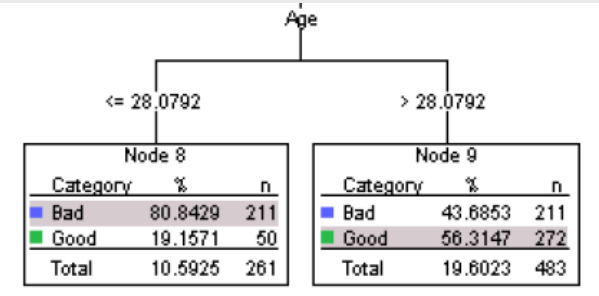
- Kakršen koli merski nivo: nominalni, ordinalni ali številski
- Če so X številski, jih je potrebno razrezati v kategorije (število kategorij lahko določi uporabnik ali algoritem)
- Pri Y se ena (npr. tvegani posojilojemalci) ali več kategorij lahko razglasi za "osnovne" (zanimive)
- Manjkajoči podatki Y: lahko so posebna kategorija ali pa se "*casewise*" izločijo
- Manjkajoči X: lahko se zamenjajo z *nadomestki* (*surrogates*), tj. drugimi prediktorji, ki so z manjkajočim X v visoki povezavi.

Metode izgradnje

- Številne, večinoma izvirajo iz *strojnega učenja*
- Način (kriteriji) izgradnje drevesa je določen z metodo - SPSS: CHAID, Izčrpni (Exhaustive) CHAID, C(A)RT, QUEST.
- Razraščanje drevesa je omejeno, npr. z:
  - do določene globine,
  - dokler je število enot v listih zadostno,
  - dokler so razlike med skupinami stat. značilne
- Pri izgradnji se kategorije X:
  - združujejo, da se doseže bolj enostavno drevo
  - razdružujejo, da je predikcija boljša.

… metode izgradnje

- CHAID (*Chi-squared Automatic Interaction Detection*): na vsakem koraku se izbere tisti prediktor, ki ima najmočnejšo interakcijo z Y. Za združevanje kategorij in razdruževanje vozlov uporablja χ^2 (Pearson ali LR).
- Izčrpni (Exhaustive) CHAID: analizira vse možne delitve (*splits*) vrednosti prediktorjev.



credit_rating * age Crosstabulation

Count

		age		
		<=28	>28	Total
credit_rating	bad	211	211	422
	good	50	272	322
Total		261	483	744

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	95,299[a]	1	,000		
Continuity Correction[b]	93,791	1	,000		
Likelihood Ratio	101,073	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	95,171	1	,000		
N of Valid Cases	744				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 112,96.

b. Computed only for a 2x2 table

... metode izgradnje

- CRT (Classification and Regression Trees). Enote razdeli v skupine, ki so čim bolj homogene glede na Y (RSS, Gini, ...). V listih (neobrezanega) drevesa imajo vse enote enak Y.
- QUEST (Quick, Unbiased, Efficient Statistical Tree). Hitra, učinkovita metoda. Nepristranska tudi pri prediktorjih z mnogimi kategorijami. Zahteva nominalno Y.

... metode izgradnje

	CHAID*	CRT	QUEST
Chi-square-based**	X		
Surrogate independent (predictor) variables		X	X
Tree pruning		X	X
Multiway node splitting	X		
Binary node splitting		X	X
Influence variables	X	X	
Prior probabilities		X	X
Misclassification costs	X	X	X
Fast calculation	X		X

*Includes Exhaustive CHAID.

**QUEST also uses a chi-square measure for nominal independent variables.

Preverjanje

- Model se preveri na novih enotah ali obstoječih (v podatkovni tabeli)
- Podatkovna tabela se (naključno) razdeli v dva – ne nujno enako velika – dela (*split-sample validation*):
  - *učni*: uporabi se za izgradnjo drevesa
  - *kontrolni*: uporabi se za preverjanje točnosti klasifikacije
- Navzkrižno preverjanje (*cross-validation*) – podatki se razdelijo v več plasti (*folds*), ki se zaporedoma izključujejo iz celotnih podatkov. Na vsakem reduciranem nizu enot se oceni *tveganje/ceno za napačno razvrstitev* (*misclassification risk/cost*)
  - poseben primer je *leave-one-out*.

Obrezovanje drevesa (*prunning*)

- Preveč razraščeno drevo – nesmiselna pravila
- Rezanje toliko časa, da ima drevo še sprejemljivo *tveganje* (*risk value*) za napačno klasifikacijo
  - tveganje se navadno izraža z deležem vseh učnih enot, ki so razvrščene narobe (*crossvalidation error*). Nekatere metode omogočajo tudi izračun standardne napake tveganja
- Možno pri CRT in QUEST.

Cena napačne razvrstitve

- *Misclassification Costs*
- Včasih je “cena” za napačno razvrstitev različna po kategorijah Y, npr.
  - strošek za banko je večji, če dodeli posojilo nekomu, ki ga ne more odplačevati, kot je izgubljeni dobiček, če ga ne odobri rednemu plačniku.
  - cena za napačno razglasitev zdrave osebe za bolnika je tipično nižja, kot če bolnika razglasimo za zdravega.
  - sprejem napačnega kandidata v službo je lahko dražji od zavrnitve dobrega.
- Določi uporabnik (glede na pretekle izkušnje).

Vnaprejšnje verjetnosti

- (*prior probabilities*)
- verjetnost pripadnosti vsaki od kategorij Y, preden poznamo X-e
- določi se lahko:
  - enaka za vse kategorije
  - enaka deležu kategorije v podatkih (učni del)
  - uporabniško določena
- vpliva na izgradnjo drevesa (samo v CRT in QUEST).

Softver

- Specializiran, npr. CART
- Splošen – statistični:
  - SPSS: modul *(IBM) SPSS Decission Trees*
  - SAS (*Enterprise Miner*), Statistica (*C&RT*),
  - R – knjižnice: *tree*, *rpart*, *party*, *randomForest*, ... (v seznamu išči tree in partition(ing)) – hiter pregled.

Priporočeno branje

- Breiman, Friedman, Olhsen, Stone: Classification and Regression Trees. Chapman&Hall, 1984.
- SPSS Decsion Trees 17.0
- JJ Faraway: Extending the Linear Model with R. Chapman&Hall, 2006 (pogl. 13, str. 253-268).