Elektrotehniški vestnik 64(5): 277±283, 1997
Electrotechnical Review, Ljubljana, Slovenija

Razširitev algoritma ReliefF

Peter Peer, Boštjan Čargo, Igor Kononenko

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: peter.peer@ananda.fri.uni-lj.si

Povzetek. Ko gradimo odločitveno drevo, moramo izmed atributov primerov, ki opisujejo učne primere, izbrati najpomembnejšega. ReliefF je eden izmed boljših algoritmov za ocenjevanje atributov, saj ni kratkoviden, omogoča pa nam ocenjevanje atributov na realnih problemih (obravnava šumne in neznane vrednosti in večrazredne probleme). Algoritem ReliefF smo popravili tako, da se je istočasno računala kvaliteta atributov za različno število bližnjih sosedov. Po opravljenih testiranjih na 20 standardnih domenah se izkaže, da so odločitvena drevesa, dobljena z algoritmom ReliefF [5], približno enako uspešna tistim, ki jih je zgradil novi algoritem ReliefFS, če primerjamo rezultate po klasiÆkacijski točnosti na testnih primerih. Vendar pa se je število listov v odločitvenih drevesih kar v nekaj primerih občutno zmanjšalo. Še posebno pride ta značilnost do izraza pri medicinskih domenah. Testiranje je zajemalo neporezana drevesa, drevesa, porezana z metodo MDL1+, in drevesa, porezana z metodo m-ocene verjetnosti. Odločitvena drevesa so bila zgrajena s sistemom Asistent-R [4] oziroma Asistent-RS. Da je razširjeni algoritem ReliefFS boljši od algoritma ReliefF, smo pokazali tudi na težkem umetnem klasiÆkacijskem problemu.

Ključne besede: strojno učenje, umetna inteligenca, odločitveno drevo, ocenjevanje atributov, ReliefF, klasiÆkacijska točnost

Extension of ReliefF

Extended abstract. When building a decision tree, we have to select the most important attribute among all attributes of instances, which form the training data set. ReliefF [4,5] is an advanced algorithm for feature selection, as it is not ™near-sighted" and can be used in real-world domains (improves certainty of estimates, deals with noisy and missing data and solves multi-class problems). We extended ReliefF in such a way, that feature qualities for different number of near neighbours are calculated simultaneously (6). After testing over 20 standard domains (Table 1, Figures 2,3), it was found, that decision trees obtained with ReliefF algorithm [5] are approximately equaly successful as those built by ReliefFS algorithm, when comparing only classiÆcation accuracy on training instances. However, another characteristic is superior: the number of leaves in decision trees is reduced heavily in many cases. This happens especially in medical domains. Our tests included unpruned trees, trees pruned with MDL1+ method [8], and Æ-nally, trees pruned with m-probability estimate [1].

Decision trees were built with Assistant-R [4] and Assistant-RS, respectively.

Extended algorithm ReliefFS is better than algorithm ReliefF. The above statement is conÆrmed also by testing results on an artiÆcially composed (synthetic) domain PCMS.

Key words: machine learning, artiÆcial intelligence, decision tree, feature selection, ReliefF, classiÆcation accuracy

1 Uvod

Strojno učenje iz primerov spada med najbolj intenzivna raziskovalna in aplikativna področja umetne inteligence. Ukvarja se z vprašanjem, kako iz podatkov, zbranih o določeni problemski domeni, izvleči obstoječe zakonitosti. Razvila se je cela vrsta metod, med katere sodi tudi induktivno učenje. Uvrščamo ga med učenje, ki temelji na podobnosti. Rezultat induktivnega učenja je formula, pravilo, teorija ali opis koncepta v kvalitativnem, logičnem formalizmu, ki je človeku dostopen in razumljiv. V atributnem strojnem učenju so posamezni primeri opisani z njihovimi lastnostmi (atributi), vsak primer pa sodi v natanko določen razred. S pomočjo induciranih pravil, ki opisujejo določen koncept, je upoštevajoč vrednosti atributov mogoče določiti razred poljubnega primera.

V množici atributov, ki opisujejo nek primer, prav gotovo niso vsi enako pomembni za predstavitev nekega koncepta, nekateri so celo nepomembni. Na zdravstveno stanje pacienta skoraj gotovo ne vpliva barva oči, krvne slike pa najbrž ne smemo zanemariti. Ocenjevanje kvalitete atributov je nasploh eden ključnih ciljev strojnega učenja. Prav pri gradnji odločitvenih dreves, ki so ena od oblik pravil, ki jih induciramo z induktivnim učenjem, je najpomembnejša izbira atributa v nekem vozlišču.

Odločitveno drevo ima za nalogo določiti razred, ki mu pripada opazovani primer. Ko gradimo drevo, moramo izmed atributov primerov, ki predstavljajo učno množico, izbrati najpomembnejšega. Najpomembnejši atribut je tisti, ki razdeli množico primerov, tako da so podmnožice čimčistejše. Za določanje pomembnosti obstaja veliko mer, nekatere izmed njih so: informativnost,

Prejet 10. november, 1997
Odobren 26. november, 1997

Gini indeks, mera razdalje, J-ocena, ustreznost, Relief, ReliefF, ocena χ^2. Večina naštetih metod ne daje dobrih ocen, če obstaja med atributi določena odvisnost. Primer t.i ™kratkovidne" ocene je informativnost, ki izbere tisti atribut, ki najbolj zmanjša entropijo razredov. Algoritem Relief je primer boljšega ocenjevanja, saj odpravlja problem obravnavanja odvisnih atributov.

Leta 1992 sta Kira in Rendell predstavila svoj kriterij za izbiro (ocenjevanje kvalitete) atributa, ki sta ga poimenovala Relief. Algoritem predpostavlja dvorazredne klasiÆkacijske probleme, bistven pomen pa ima izbira bližnjih sosedov glede na pomembne atribute [4]. Algoritem ReliefF [5], razširitev osnovnega postopka Relief, omogoča uporabo algoritma v praksi na realnih problemih, saj poveča zanesljivost ocen, obravnava šumne in neznane vrednosti in večrazredne probleme.

Algoritem ReliefF smo popravili tako, da se je istočasno računala kvaliteta atributov za različno število bližnjih sosedov. Končna ocena kvalitete atributa je postala največja izmed vseh tako dobljenih ocen. Osnovna ideja je v tem, da število bližnjih sosedov lokalno prilagodimo podprostoru, ki mu ustreza trenutno izbrani primer [8].

Končno smo algoritma ReliefF in ReliefFS testirali na standardnih domenah in na umetno sestavljeni domeni, ki je pokazala pomanjkljivost prvega in učinkovitost slednjega.

Vsa odločitvena drevesa so bila zgrajena po sistemu Asistent-R [4] oziroma Asistent-RS. Asistent-R predstavlja avtomatsko učenje odločitvenih dreves s sistemom za ocenjevanje atributov ReliefF, Asistent-RS pa avtomatsko učenje odločitvenih dreves s sistemom za ocenjevanje atributov ReliefFS.

V realnih domenah so podatki često šumni. Osnovni algoritem zgrajeno drevo tipično preveč prilagodi učnim primerom, kar se kaže v večjih drevesih in nezanesljivi klasiÆkacijinovih primerov. Rezanje dreves, katerega namen je odstranitev nezanesljivih poddreves, zmanjša kompleksnost dreves in hkrati poveča klasiÆkacijsko točnost na neodvisnem vzorcu. Poznamo dve glavni vrsti rezanja:

- predhodno rezanje in
- naknadno rezanje.

Metodi rezanja m-ocena verjetnosti in MDL1+, ki smo jih uporabili pri testiranju, spadata v skupino naknadnega rezanja.

Metoda m-ocena verjetnosti [1] pravi, da k oceni verjetnosti prispevata svoj delež relativna frekvenca n/N in apriorna verjetnost p_a. Ko je m po absolutni vrednosti velik, se bolj kot na izide opravljenih poskusov zanašamo na apriorno verjetnost (če je $m = \infty$, rezultatov sploh ne upoštevamo). Ko pa je m blizu 0, se bolj kot na naše predznanje zanesemo na izide poskusov (če je $m = 0$, vpliv apriorne verjetnosti izgine). Metoda poreže drevo v vozlišču takrat, ko je napaka porezanega drevesa manjša ali enaka napaki neporezanega drevesa. Ta princip rezanja imenujemo rezanje z minimalno napako.

Princip najkrajše dolžine opisa (Minimum Description Length - MDL) pravi, da je najbolj verjetna razlaga nekih podatkov tista teorija izmed vseh tistih, ki reproducirajo te podatke, katere opis je najkrajši. Metoda MDL1 [3] upošteva samo dolžino opisa listov in v splošnem slabo reže. Popravek, ki za vsako notranje vozlišče prišteje vrednost 1, nam da metodo MDL1+ [8], ki pa daje zelo dobre rezultate. Metoda poreže drevo v vozlišču takrat, ko je opis lista krajši od opisa poddrevesa.

2 Ocenjevanje atributov

2.1 Učni in testni primeri

Da lahko natančno določimo razred opazovanega primera, morajo izbrani atributi, ki jih je lahko malo ali veliko, razbiti množico primerov na (popolnoma) čiste podmnožice - razrede. Problem izbire je v tem, da ne vemo, kateri atribut nam največ pove o pripadnosti posameznim razredom.

Domena, na kateri izvajamo testiranje uspešnosti zgrajenega (in kasneje porezanega) odločitvenega drevesa, je množica primerov. Vsi primeri imajo enako število atributov in pripadajo nekemu razredu. Da bi lahko simulirali klasiÆkacijo na neznanih primerih, razbijemo domeno na dva dela: na učne in na testne primere. Učni primeri so tisti, na osnovi katerih algoritem zgradi odločitveno drevo. Testni primeri nam predstavljajo še neznane primere in so merodajni pri določanju klasiÆkacijske točnosti tako zgrajenega drevesa. Da bi dobili bolj verodostojne rezultate, postopek razbitja in testiranja ponavljamo. Razbitja so naključna. Razmerje med številom učnih in številom testnih primerov je 70:30 odstotkov.

2.2 ReliefF

ReliefF [4,5], razširitev osnovnega postopka Relief, omogoča uporabo algoritma v praksi na realnih domenah. Med različnimi algoritmi razvitimi iz originalnega algoritma se je prav ReliefF izkazal kot najbolj celovit in učinkovit. Smiselno je deÆniran prehod iz dvorazrednih na večrazredne probleme, rešuje pa tudi problem manjkajočih vrednosti atributov, ki je še posebej pereč v medicinskih domenah (često se namreč zgodi, da na pacientu ne moremo izvesti vseh potrebnih preiskav).

Algoritem ReliefF, ki je na sliki 1 med dvema komentarjema, oceni utež $\boldsymbol{W}[a]$ atributa X_a kot razliko verjetnosti:

$$\boldsymbol{W}[a] = P(\text{različna vrednost atributa}|\text{bližnji pogrešek}) - P(\text{različna vrednost atributa}|\text{bližnji zadetek}) \quad (1)$$

Postopek ReliefF uporablja zaradi zanesljivejših ocen verjetnosti namesto enega samega bližnjega zadetka oziroma pogreška primera I, k bližnjih zadetkov $\mathcal{Z}^+$ oziroma pogreškov $\mathcal{Z}^-$ ter nato upošteva povprečje prispevkov posameznih zadetkov. Pri večrazrednih problemih pa ne išče le k bližnjih pogreškov, temveč upošteva k bližnjih pogreškov $\mathcal{Z}_C^-$ iz vsakega različnega razreda.

Ime dom.	ReliefF						ReliefFS					
	Pred rez.		M-2		MDL1+		Pred rez.		M-2		MDL1+	
	Klas. toč.	Št. lis.	Klas. toč.	Št. lis.	Klas. toč.	Št. lis.	Klas. toč.	Št. lis.	Klas. toč.	Št. lis.	Klas. toč.	Št. lis.
ASIS	78.0	29.0	82.0	18.1	83.3	11.1	78.0	28.5	80.6	19.8	82.2	11.3
BAYS	72.0	25.1	70.5	16.0	70.8	8.2	68.1	19.9	69.9	12.4	69.9	7.0
P2AR	94.7	8.4	95.2	4.6	96.5	4.0	94.7	8.2	95.2	4.6	96.5	4.0
P3AR	89.8	13.5	93.9	8.7	95.0	8.2	90.2	13.4	93.9	8.7	95.0	8.2
P400	54.5	5.8	55.1	4.6	55.1	4.6	57.1	6.3	57.4	4.7	57.4	4.1
BO1L	78.0	23.3	78.0	9.3	78.0	9.3	67.7	8.8	67.7	6.3	67.7	6.7
VO1E	95.0	5.5	95.1	5.1	95.1	4.6	95.0	6.8	95.2	6.0	95.0	5.1
SEGM	87.1	33.5	87.2	26.1	86.6	19.0	84.5	13.6	84.4	12.2	84.3	12.3
IRIS	96.8	4.9	96.5	4.2	97.2	3.2	97.5	3.0	97.5	3.0	97.5	3.0
K1RK	97.8	8.4	97.8	8.4	97.8	8.4	97.8	8.4	97.8	8.4	97.8	8.4
K2RK	66.9	1.0	66.9	1.0	66.9	1.0	67.2	3.1	67.2	2.6	67.2	2.6
ANKE	66.3	36.5	66.3	32.1	67.3	10.2	66.0	22.0	65.8	17.7	67.5	7.6
BREA	78.6	18.4	78.6	9.6	80.0	3.5	78.8	9.4	78.8	6.0	78.8	3.1
DIAB	72.4	23.5	72.1	14.8	72.9	13.9	72.9	8.6	73.2	6.6	72.7	7.3
HEAR	76.2	27.4	77.8	17.1	76.7	11.5	73.6	13.3	73.6	6.2	73.6	6.3
HEPA	81.1	8.2	81.1	7.8	75.6	2.0	77.2	10.6	77.7	8.5	75.6	1.7
LYMP	77.5	17.6	77.5	16.4	75.2	7.1	74.4	18.4	76.2	14.6	76.5	6.8
ST2G	74.9	8.2	74.9	7.7	65.9	3.2	74.7	3.9	74.7	2.9	63.8	1.0
ST5G	58.9	9.5	58.9	7.0	62.1	5.3	60.3	8.1	61.6	5.7	66.4	3.8
E1PI	77.7	19.7	77.7	13.7	77.7	10.9	77.3	12.1	77.3	10.9	77.1	10.1

Tabela 1. Rezultati testa na standardnih domenah

Table 1. Test results over standard domains

Povprečja posameznih prispevkov nato obteži z apriorno verjetnostjo razreda:

$$\begin{aligned} \boldsymbol{W}[a] = \quad & \boldsymbol{W}[a] - \left[\tfrac{1}{k}\sum_{z^+\in\mathcal{Z}^+}\text{diff}(I,z^+,a)\right]^2 + \\ & [\sum_{C\neq razred(I)}[\,P(C)\times \\ & \left(\tfrac{1}{k}\sum_{z^-\in\mathcal{Z}_C^-}\text{diff}(I,z^-,a)\right)\,]\,]^2 \end{aligned} \tag{2}$$

Algoritem lahko s takšnim pristopom oceni, kako atribut razlikuje med poljubnima razredoma in ne le med najbližjima.

Razliko med vrednostmi atributov dvah primerov I_1 in I_2 določa funkcija diff, ki je posebej definirana za diskterne in zvezne atribute. Za diskretne atribute a se glasi

$$\text{diff}(I_1,I_2,a) = \begin{cases} 0, & vrednost(I_{1,a}) = vrednost(I_{2,a}) \\ 1, & vrednost(I_{1,a}) \neq vrednost(I_{2,a}) \end{cases}, \tag{3}$$

za zvezne pa

$$\text{diff}(I_1,I_2,a) = \frac{vrednost(I_{1,a}) - vrednost(I_{2,a})}{n}, \tag{4}$$

kjer je n normalizacijska konstanta, s katero normaliziramo vrednost dunkcije diff na interval $[0,1]$.

Oglejmo si še kako algoritem izračuna verjetnost, da imata dva primera I_1 in I_2 različni vrednosti a-tega atributa v primeru manjkajočih vrednosti:

- če ima en primer (npr. I_1) manjkajočo vrednost:

$$\text{diff}(I_1,I_2,a) = 1 - P(vrednost(I_2,a)|razred(I_1)), \tag{5}$$

- če sta neznani vrednosti obeh primerov:

$$\begin{aligned} \text{diff}(I_1,I_2,a) = 1 - \sum_{v=1}^{V_a}(& P(v|razred(I_1))\times \\ & P(v|razred(I_2))\,), \end{aligned} \tag{6}$$

kjer je V_a število vrednosti a-tega atributa.

Pogojne verjetnosti izračunamo iz učne množice po enem izmed načinov za ocenjevanje verjetnosti.

Prednosti algoritma ReliefF se pokažejo predvsem pri atributih z močno interakcijo. V primeru, ko so atributi neodvisni, se izkaže, da je ocena algoritma ReliefF ob upoštevanju velikega števila bližnjih zadetkov oziroma pogreškov (pri ocenjevanju v bistvu upoštevamo celotno učno množico) v močni korelaciji z Gini-indeksom, ki je v močni korelaciji z informativnostjo.

2.3 Razširitev algoritma ReliefF

Na velikost bližnje soseščine lahko vplivamo s spreminjanjem parametra k. Parameter k namreč določa število

Ime domene	Število razredov	Število atributov	Število vrednosti/atribut	Število primerov	Del. večinskega razreda (%)	Entropija razreda (bit)
ASIS	2	11	2.0	200	57	0.99
BAYS	2	10	2.0	200	56	0.99
P2AR	2	12	2.0	200	54	0.99
P3AR	2	13	2.0	200	54	0.99
P400	2	14	2.0	400	50	1.00
BO1L	2	6	2.0	640	67	0.91
VO1E	2	16	2.0	435	61	0.96
SEGM	7	19	8.3	2310	14	2.81
IRIS	3	4	6.0	150	33	1.59
K1RK	2	18	2.0	1000	67	0.92
K2RK	2	6	8.0	1000	67	0.92
ANKE	6	32	9.1	355	66	1.73
BREA	2	10	2.7	288	80	0.73
DIAB	2	8	8.8	768	65	0.93
HEAR	2	13	5.0	270	56	0.99
HEPA	2	19	3.8	155	79	0.74
LYMP	4	18	3.3	148	55	1.28
ST2G	2	19	4.5	270	65	0.94
ST5G	5	19	4.5	270	65	1.63
E1PI	2	18	16.6	500	50	1.00

Tabela 2. Osnovni podatki o domenah

Table 2. Basic description of data sets

bližnjih sosedov, ki jih poiščemo pred ocenjevanjem pomembnosti atributov. ReliefF smo zato razširiti tako, da se istočasno računa kvaliteta atributov za različne vrednosti parametra k, to pa pomeni, da k vpliva na izbiro (glede na kvaliteto) najbolje ocenjenega atributa. Končna ocena kvalitete atributa a je potem:

$$W[a] = \max_k W[a, k]. \quad (7)$$

Popravili smo torej funkcijo, ki je zadolžena za izračun kvalitete atributov. **Namesto za konstanten parameter k je morala funkcija med sabo primerjati kvalitete atributov preko celotne zaloge vrednosti za k.** Slika 1 prikazuje popravljen algoritem ReliefF.

3 Primerjava ReliefF - ReliefFS

3.1 Standardne domene

Pri testiranju smo uporabili naslednje standardne domene:

- skupina umetnih domen: ASIS, BAYS, P2AR, P3AR, P400, BO1L, K1RK, K2RK
- skupina medicinskih domen: ANKE, BREA, DIAB, HEAR, HEPA, LYMP, ST2G, ST5G
- skupina ostalih realnih domen: E1PI, VO1E, SEGM, IRIS

Tabela 2 podaja osnovne podatke zgornjih domen.

Rezultati testiranj na standardnih domenah so pokazali, da se pri večini domen klasifikacijska točnost* odločitvenih dreves, dobljenih z novim algoritmom, ni bistveno (signifikantno) spremenila. Vzrok za to je predvsem dejstvo, da so domene imele tudi po več sto primerov, kar je pri vrednosti $k = 10$ botrovalo uspehu algoritma ReliefF. Vendar pa je opazna tudi ena značilnost, namreč število listov v odločitvenih drevesih se kar v nekaj primerih občutno zmanjša. Še posebno pride to do izraza pri medicinskih domenah. Rezultati so razvidni iz tabele 1.

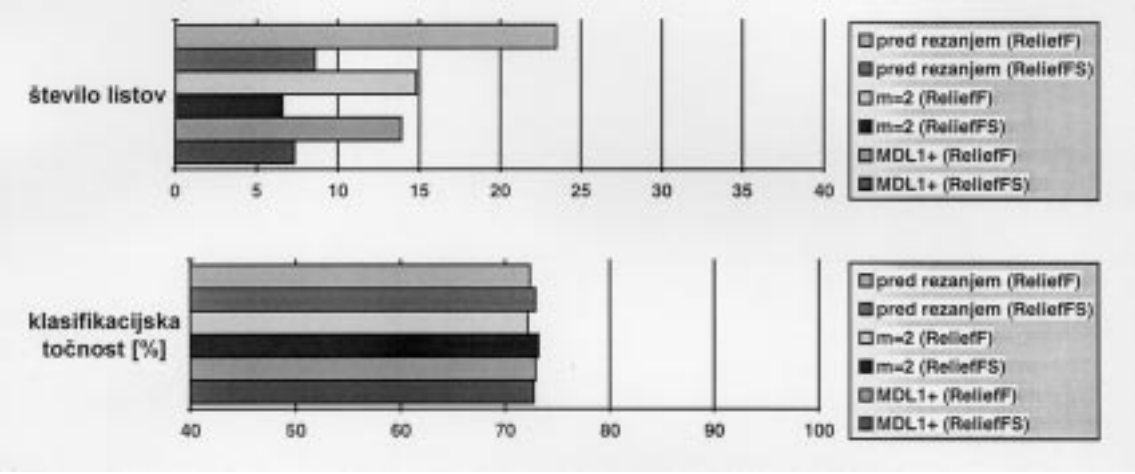
Slika 2 prikazuje rezultate testa nad medicinsko domeno DIAB, ki predstavlja opis razširjenosti diabetesa pri Indijankah iz plemena Pima. Iz slike lahko razberemo, da se število listov v odločitvenem drevesu ob uporabi novega algoritma občutno zmanjša, klasifikacijska točnost pa se ne spremeni signifikantno. Torej dobimo manjša odločitvena drevesa.

Enak zaključek nam ponuja tudi slika 3, ki pa prikazuje rezultate testa nad domeno SEGM iz področja računalniškega vida. Ta domena predstavlja problem segmentacije slik.

*Klasifikacijska točnost podaja odstotek pravilno klasificiranih testnih primerov.

Vhodni podatki:
množica učnih primerov $\mathcal{L}$, parameter m_r, ki podaja število naključno izbranih primerov učne množice
Izhod:
vektor pomembnosti atributov $\boldsymbol{MaxPomembnost}$
Algoritem:
za vse atribute: $MaxPomembnost[a] = 0$
za vsak k:
// ReliefF - začetek
razdeli $\mathcal{L}$ na N množic $\mathcal{L}_1$, $\mathcal{L}_2$, ..., $\mathcal{L}_N$, kjer je N število razredov
za vse atribute: $W[a] = 0$
za i je 1 **do** m_r **ponovi**
naključno izberi primer $I \in \mathcal{L}$
za c je 1 **do** N
če c enak $razred(I)$ **potem**
med primeri $\mathcal{L}_c$ poišči k najbližjih zadetkov $\mathcal{Z}^+$
ažuriraj utež (W,I,$\mathcal{Z}^+$)
sicer
med primeri $\mathcal{L}_c$ poišči k najbližjih pogreškov $\mathcal{Z}_c^-$
ažuriraj utež (W,I,$\mathcal{Z}^-$,$P(C)$)
$\boldsymbol{Pomembnost} = \frac{1}{m_r} \cdot \boldsymbol{W}$
// ReliefF - konec
za a je 0 **do** število atributov domene
če $Pomembnost[a] > MaxPomembnost[a]$ **potem**
$MaxPomembnost[a] = Pomembnost[a]$
vrni $\boldsymbol{MaxPomembnost}$

Slika 1. Algoritem ReliefFS Figure 1. ReliefFS algorithm



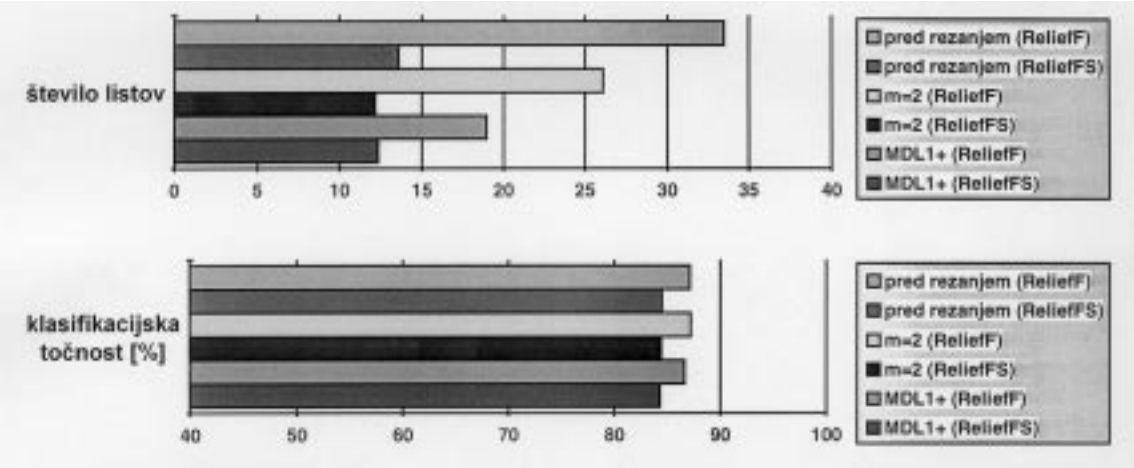
Slika 2. Rezultati testiranja obeh algoritmov na medicinski domeni DIAB
Figure 2. Test results of both algorithms over medical domain DIAB

3.2 Nova umetna domena

V tem razdelku je opisana primerjava algoritma na *novi* umetni domeni. Sestavili smo umetno domeno, katere naloga je bila, da pokaže prednost novega algoritma (ReliefFS) in slabost starega (ReliefF). Domeno smo sestavili tako, da smo ™prevarali" star algoritem, tako da je izbiral napačne atribute, nov algoritem pa prave. To se je odražalo v višji klasiÆkacijski točnosti. Velikost prednosti v klasiÆkacijski točnosti pa je pravzaprav nepomembna, bistveno je, da so se pri primerjavi pokazale statistično pomembne razlike.

Kdaj bo star algoritem izbiral napačne atribute (glede na k)? - To se bo zgodilo pri majhni učni množici, kjer imamo močno interakcijo med atributi; število primerov, ki pripadajo isti soseščini, mora biti manjše od k (k = 10 pri starem algoritmu).

Kakšna je sploh domena, ki je bila testirana? Kot osnova je vzeta domena P2AR, ki predstavlja problem parnosti drugega reda. Pri tej domeni sta najpomembnejša prva dva atributa. Očitno je, da bo nov algoritem



Slika 3. Rezultati testiranja obeh algoritmov nad domeno SEGM iz področja računalniškega vida
Figure 3. Test results of both algorithms in SEGM domain from the computer vision area

deloval ™bolje" pri majhni množici primerov v domeni, saj to narekuje dejstvo, da moramo v domenah z močno interakcijo med atributi paziti, da parameter k ne bo večji od števila primerov, ki pripadajo isti soseščini! Na podlagi tega dejstva je iz domene P2AR nastala domena PCMS, ki vsebuje le 20 primerov v domeni; domena P2AR vsebuje 200 primerov v domeni. Domena PCMS pa se od domene P2AR razlikuje še po dveh lastnostih:

- namesto 12 atributov ima 14 atributov in
- odvisnost med atributi je še večja in bolj opazna.

Prej sta bila medsebojno odvisna prvi in drugi atribut, zdaj pa so med seboj odvisni prvi štirje atributi. Parnost pravi, da če je vsota vrednosti pomembnih atributov liha, potem pripada primer razredu 1, sicer pa razredu 2. V domeni PCMS pa naj bi bila odvisnost med atributi še bolj opazna zato, ker ob istih vrednostih prvih štirih atributov primer pripada razredu 2, sicer pa razredu 1. Domena je še dodatno pokvarjena z 2.5% šuma v vrednostih razreda.

Tabela 3 prikazuje povprečno klasiÆkacijsko točnost na domeni PCMS pri 10 ponovitvah poskusa. KlasiÆkacijska točnost podaja odstotek pravilno klasiÆciranih neznanih, torej testnih, primerov.

ReliefF		ReliefFS	
M-2	MDL1+	M-2	MDL1+
52.0%	50.6%	68.8%	66.6%

Tabela 3. Povprečna klasiÆkacijska točnost na domeni PCMS
Table 3. Average classiÆcation accuracy in PCMS domain

Za testiranje pomembnosti razlik smo uporabili dvosmerni parni t-test. Vrednost eksperimentalnega t:

- Metoda M-2: $|t|$=3.28 $\Rightarrow$ 99% stopnja zaupanja, da razlika v doseženi točnosti med algoritmoma ni naključna.
- Metoda MDL1+: $|t|$=2.82 $\Rightarrow$ 98% stopnja zaupanja, da razlika v doseženi točnosti med algoritmoma ni naključna.

4 Sklep

Rezultati naših testiranj so pokazali, da se pri večini domen klasiÆkacijska točnost dreves, dobljenih z novim algoritmom, ni bistveno spremenila (tabela 1). Vzrok za to so razmeroma učinkovita odločitvena drevesa, ki jih je zgradil že algoritem ReliefF. Prav tako so nekatere vrednosti, ki jih v ReliefFS prirejamo parametru k, večje od števila primerov bližnje soseščine in celo večje od števila vseh primerov domene. Seveda moramo upoštevati, da to velja pretežno za realne domene z večinoma neodvisnimi atributi primerov. Umetna domena, ki smo jo sestavili sami (razširjen problem parnosti pri domeni P2AR), in v kateri vlada med atributi dokaj močna odvisnost, kaže drugačno sliko. Razlika v klasiÆkacijski točnosti, ki jo dajeta ReliefF in ReliefFS, je statistično signiÆkantna.

S poskusi smo potrdili hipotezo, da je ReliefFS koristna nadgraditev algoritma ReliefF.

5 Literatura

[1] B. Cestnik, I. Bratko, On Estimating Probabilities in Tree Prunning, *Proc. European Working Session on Learning EWSL-91*, Porto, Portugal, April 1991.

[2] I. Kononenko, On Biases in Estimating Multi-Valued Attributes, *Proc. Int. Joint Conf. On ArtiÆcial Intelligence IJCAi-95*, Montreal, Canada, August 1995, pp. 1034-1040.

[3] M. Mehta., J. Rissanen, R. Agrawal, MDL-based Decision Tree Prunning, *Proc. 1st Int. Conf. On Knowledge discovery in databases and Data mining KDD-95*, Montreal, Canada, August 1995, pp. 216-221.

[4] E. Šimec, Avtomatsko učenje odločitvenih dreves s sistemom za ocenjevanje atributov ReliefF, Diplomska naloga, *Univerza v Ljubljani*, FER, Ljubljana.

[5] I. Kononenko, Estimating Attributes: Analysis and Extensions of RELIEF, *Proc. ECML-94*, Catania, Italy, april 1994, pp. 171-182.

[6] J. Kampuš, A. Počič, J. Resnik, Rezanje odločitvenih dreves po principu MDL, *seminarska naloga pri predmetu HPUI*, FRI, Ljubljana, maj 1996.

[7] P. Peer, B. Čargo, T. Mele, Rezanje odločitvenih dreves po principu MDL in razširitev algoritma ReliefF, *seminarska naloga pri predmetu UISP*, FRI, Ljubljana, september 1997.

[8] I. Kononenko, *Napotki za izdelavo seminarske naloge*, FRI, Ljubljana, 1996.

[9] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF, *Applied Intelligence 7*, 1997, pp. 39-55.

Peter Peer in **Boštjan Čargo** sta študenta petega letnika na Fakulteti za računalništvo in informatiko, Univerze v Ljubljani.

Igor Kononenko je izredni profesor na Fakulteti za računalništvo in informatiko v Ljubljani. Raziskovalno se ukvarja z razvojem algoritmov strojnega učenja. Je (so)avtor sedmih učbenikov (ena knjiga izšla v tujini) ter 90 objavljenih člankov in referatov (70 v tujini).