INTERAKTIVNA INTERAKCIJSKA ANALIZA

Aleks Jakulin, Gregor Leban
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Tržaška 25, SI-1001 Ljubljana, Slovenia
Tel: +386 1 4768 813; fax: +386 1 1 4768 386
e-mail: jakulin@acm.org

POVZETEK

Interakcije lahko razumemo kot korelacije, ki obsegajo več kot le dva atributa. Neka skupina atributov je med seboj v interakciji, če njihovih medsebojnih povezanosti ne moremo popolnoma razumeti, ne da bi jih vse opazovali hkrati. Interakcije so zakonitosti skupin več atributov.

V tem članku merimo pomembnost interakcije s postopki, ki temeljijo na Shannonovi entropiji kot pojmu negotovosti, ki je bolj splošen od koncepta statistične variance. Cilj interakcijske analize je analitiku predstaviti interakcije grafično z več tipi diagramov. S tem namenom smo izdelali orodja, ki omogočajo interaktivno preučevanje podatkov in nudijo pomoč pri iskanju zanimivih pogledov na podatke. Interakcije prinašajo tudi nov pogled na nekatere težave postopkov strojnega učenja.

1 UVOD

Ko poskušamo ljudje razumeti podatke, jih ne obravnavamo v celoti. Raje jih razbijemo na manjše koščke, ki so bolj obvladljivi. To deljenje problemov na podprobleme je osnova večine postopkov strojnega učenja. Čeprav je redukcionističen, deluje.

A obstajajo delčki znanja in vzorci v naravi, ki izginejo, če jih poskušamo razrezati. Moramo jih obravnavati kot celoto. Po drugi strani pa spet ne moremo vsega obravnavati kot celoto, saj je poenostavljanje ključno za zmožnost posploševanja.

Da bi prerezali ta gordijski vozel, vpeljimo koncept *interakcij*. Interakcije so tisti vzorci, ki jih ne moremo razumeti po koščkih, le v celoti. Problem lahko prosto razbijamo na koščke, če le ne razbijemo interakcij.

Predstavljajmo si marsovskega bankirja, ki bi rad stranke razdelil v tri razrede: goljufe, povprečneže in molzne krave. Bankir ima na voljo množico spremenljivk, ki stranko opisujejo: starost, poklic, izobrazbo, lanskoletne dohodke, letošnje dohodke in dolgove.

Bankir zaposluje več analitikov. Najraje bi predpostavil, da so vse spremenljivke med seboj neodvisne, a hkrati tudi vse povezane z razredom. Potem bi lahko vsakemu analitiku predal v študij le po eno spremenljivko. Vsak analitik je strokovnjak o odnosu med svojo spremenljivko in razredom, izkušnje pa je pridobil na velikem številu primerov, ki jih je že preučeval. Ko analitiki odhitijo s podatki, med seboj ne komunicirajo: samo na podlagi svoje spremenljivke se poskušajo odločiti, v katerem razredu je nova stranka.

Bankir čez nekaj časa skliče vse analitike in jim pove, naj glasujejo za posamezen razred. Če nek analitik čuti, da nima dovolj podatkov, mu je dovoljeno, da se vzdrži glasovanja. Bankir izbere razred, ki je dobil največ glasov. V primeru, da je takih razredov več, izbere najslabšega: vsekakor je boljše, da obravnava molzno kravo kot goljufa, kot pa da bi klečeplazil pred goljufom.

Žal sta tu dve težavi. Več analitikov lahko preučuje iste informacije. Na primer, ko enkrat poznamo strankin poklic, nam njena izobrazba ne bo povedala kaj bistveno novega. Zato bo ta plat stranke dobila preveliko težo pri glasovanju. Takim spremenljivkam pravimo, da so *soodvisne*.

Druga težava je v tem, da nam lanskoletni dohodki in letošnji dohodki ne povedo toliko, kot bi nam povedali, če bi namesto tega vedeli, kako so se dohodki spremenili. Na primer, povprečneži se lahko spreobrnejo v goljufe, če se jim dohodki na hitro zmanjšajo. Za take spremenljivke pa rečemo, da so *sodejavne*.

Interakcije so pojem, ki združuje sodejavnosti in soodvisnosti. V primeru interakcij se splača, da analitiki med seboj sodelujejo, da bi dosegli boljše rezultate. Malo bolj realistično: en analitik naj obdeluje več spremenljivk in jih združi v eni sami formuli. Na primer, dve spremenljivki o dohodkih zamenjamo z indeksom padca dohodkov, kar je nova spremenljivka, na podlagi katere analitik sprejme svojo odločitev. Tako smo se interakcije rešili.

2 INTERAKCIJE SKOZI ENTROPIJO

Entropijo je definiral Shannon [1] z namenom količinskega merjenja informacijske vsebine. Lahko jo razumemo kot posplošitev koncepta variance, kjer nismo omejeni na parametrične verjetnostne porazdelitve z enim samim lokalnim maksimumom. Naključna spremenljivka ima visoko entropijo, ko je njeno vrednost težko predvideti, nizko pa, ko je vrednost predvidljiva. Dogodek z verjetnostjo p=0,5 ima večjo entropijo kot dogodek s p=0,9. Če je možnih N dogodkov glede na neko diskretno porazdeljeno naključno spremenljivko A, verjetnost i-tega pa je p_i, v bitih merjeno entropijo računamo tako:

$$H(A) \equiv -\sum_{i=1}^{N} \log_2 p_i$$

V primeru, da imamo dve diskretno porazdeljeni slučajni spremenljivki A in B, prva z N, druga pa z M dogodki, obstaja največ NM možnih skupnih dogodkov. Njuno skupno entropijo opišemo kot:

$$H(AB) \equiv -\sum_{j=1}^{M}\sum_{i=1}^{N} \log_2 p_{i,j}$$

AB lahko razumemo kot novo slučajno spremenljivko, katere množica dogodkov je kartezični produkt množic, ki pripadata A in B posamično. Skupna entropija $H(AB)$ je manjša ali kvečjemu enaka vsoti posamičnih entropij $H(A)+H(B)$. Manjša je takrat, ko sta spremenljivki odvisni, odvisnost pa količinsko izmerimo z *medsebojno informacijo* ali *informacijskim prispevkom*:

$$I(A;B) \equiv H(A) + H(B) - H(AB)$$

Če odnos med dvema spremenljivkama razumemo kot 2-interakcijo, je medsebojna informacija dokazano učinkovita mera za količino informacije, ki je skupna obema spremenljivkama. Kot taka je poskus kvantifikacije konceptov, ki smo jih omenjali prej. McGill [2] je predlagal posplošitev koncepta medsebojne informacije na več spremenljivk, kar je za njim in neodvisno od njega bilo še večkrat predlagano [3]. Količino imenujemo *interakcijska informacija* ali *interakcijski prispevek*, za primer treh spremenljivk A,B,C je definirana na način, ki spominja na načelo vključitve in izključitve v teoriji množic:

$$I(A;B;C) \equiv I(AB;C) - I(A;C) - I(B;C)$$

Interakcijski prispevek je lahko pozitiven ali negativen. Pozitivni interakcijski prispevek odraža pozitivno interakcijo ali sodejavnost, negativni interakcijski prispevek pa negativno interakcijo ali soodvisnost. Interakcijski prispevek je dejansko vsota pozitivne *sinergije* in negativne *odvečnosti* teh treh spremenljivk, ki pa ju ne še znamo meriti neposredno. Sinergija se izkazuje kot specifična posebnost, ki jo dojamemo le takrat, ko opazujemo hkratno pojavljanje vseh treh slučajnih spremenljivk. Odvečnost pa je medsebojna informacija, ki sama ni nič drugega kot 2-interakcijski prispevek in se je prenesla v 3-interakcijski prispevek. Na primer, če imata 2-interakciji med A in B ter B in C nekaj skupnega, se to odraža kot odvečnost pri 3-interakciji.

Pri soodvisnostih velja, da nam spremenljivki A in B podata deloma iste informacije, ki nam skupaj povejo manj o C, kot bi pričakovali, če bi kar sešteli obseg informacij, ki jih podata A in B posamično. Po drugi strani nam pri sodejavnostih spremenljivk A in B, če ju obravnavamo hkrati, podata informacije o C, ki nam jih ne bi podala nobena od spremenljivk neodvisno od drugih.

2 VIZUALIZACIJA INTERAKCIJ

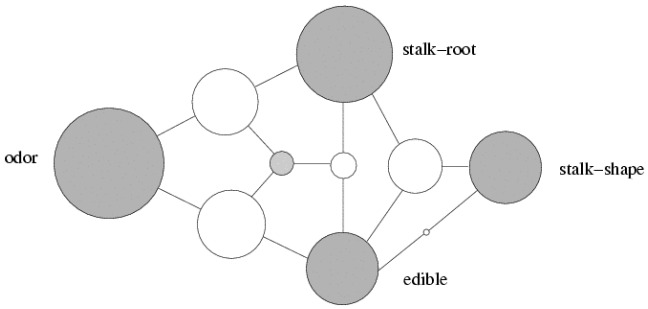
V strojnem učenju najpogosteje reševan problem je klasifikacija ali uvrščanje. Pri takem primeru nam je danih večje število *učnih primerov*, od katerih je vsak opisan z več *atributi*. Eden od atributov je *razredni* in opisuje razred, v katerega je primer uvrščen. Cilj učenja je na učnih primerih najti zakonitosti, vzorce in pravila, ki nam omogočajo, da bomo za primer, ki ga še nismo videli, s čimvečjo verjetnostjo napovedali njegov pravi razred. Ključna je torej sposobnost posploševanja.

V tako definiranem problemu nas še posebej zanimajo odnosi med dvema navadnima atributoma in razrednim. Vse tri atribute, pa tudi množice le-teh lahko modeliramo kot naključne spremenljivke, tako kot smo to opisali v prejšnjem razdelku. Do slučajne spremenljivke, ki ustreza enemu od atributov, pridemo kar tako, da preštejemo delež učnih primerov, ki ima neko konfiguracijo vrednosti atributov. Dogodek j spremenljivke A je učni primer, pri katerem ima istoimenski atribut vrednost z zaporedno številko j. Ta postopek deluje najbolje za neurejene diskretne oziroma nominalne atribute. Ni težko videti, kako postopek deluje, če obravnavamo po dva ali več atributov hkrati.

Interakcijski prispevek nam ponuja vpogled v odnose med dvema atributoma ter razredom. Vendar je to le število, ki uprabniku, ki bi želel razumeti vzorce v podatkih, ne nudi ravno najboljšega vpogleda. Zato bomo uporabili nekatere znane postopke prikazovanja podatkov iz statistične in matematične vizualizacije.

3.1 Interakcijski diagram

Entropijo, informacijski ter interakcijski prispevek merimo v bitih. Zato lahko vse te količine prikažemo z isto mero. Vsaki količini priredimo krog, katerega ploščina ustreza skupni informaciji. Krog je svetlo obarvan, če gre za sinergijo, temno pa v primeru entropije ali odvečnosti. Količina nastopa kot vozlišče v grafu, kjer je povezana s količinami, katerih odvečnost ali sinergijo predstavlja.



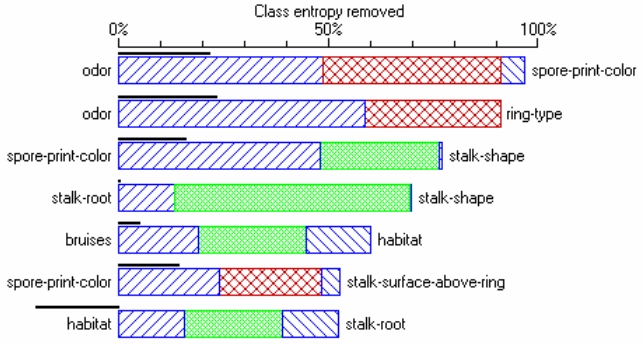
Slika 1: *Interakcijski diagram s ploščino vozlišč opiše informacijske ter entropijske vsebine med naključnimi spremenljivkami.*

Na sliki 1 je prikazanih nekaj atributov iz podatkovne zbirke o ameriških gobah. Razredni atribut je užitnost (*edible*), ki opisuje, ali je goba užitna. Njegova entropija znaša približno 1 bit, kar ustreza naši začetni negotovosti glede užitnosti gobe. Navadni atributi opisujejo vonj gobe (*odor*), obliko kocena (*stalk-shape*) ter dna kocena (*stalk-root*). Na podlagi velikosti belih krogov, ki povezujejo atribute vidimo, da ima največji informacijski prispevek k razredu atribut vonj, najmanjšega pa atribut oblike kocena.

Med atributi so zanimive povezave. Temni krožec označuje količino informacije, ki ga soodvisna atributa vonja in oblike dna kocena oba prispevata o razredu: to količino moramo odšteti od vsote njunih prispevkov, saj je to odvečnost. Po drugi strani pa obstaja zelo močna sinergija med sodejavnima *stalk-root* in *stalk-shape*, ki jo moramo prišteti k njunim posamičnim informacijskim prispevkom. Četudi v diagramu tega nismo prikazali, obstaja tudi interakcijski prispevek med vsemi štirimi atributi, ki je tu negativen. Nismo prikazali niti nekaterih manj izrazitih interakcij: vizualizacija mora iskati ravnotežje med celovitostjo in bistvom.

3.2 Grafični pregled interakcij

Četudi bi lahko interakcijski prispevek obravnavali neposredno v bitih, je pri nalogi uvrščanja bolj koristno izraziti vse informacijske količine kot deleže začetne negotovosti glede razreda. Podatek o vsakem atributu odstrani delež negotovosti glede razreda, ki ustreza njuni medsebojni informaciji. Tu soodvisnosti izrazimo kot prekrivanje (kar spominja na Vennove diagrame, ki prikazujejo preseke med množicami), sodejavnosti pa kot razmike med posamičnimi prispevki.

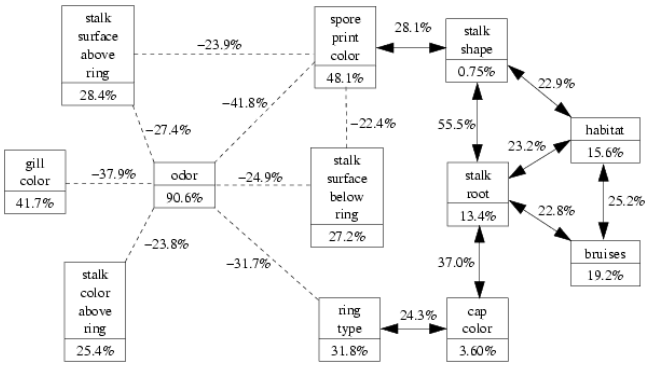


Slika 2: *Pregled interakcij nam prikaže soodvisnosti in sodejavnosti izražene glede na entropijo razreda.*

Na sliki 2 vidimo nekatere od interakcij v domeni o gobah. Najpomembnejša soodvisnost je med vonjem gobe (*odor*) in barvo trosov (*spore-print-color*). Vonj sam po sebi nam odpravi več kot 90% negotovosti, barva trosov pa malo manj kot 50%. Očitno nam oba atributa ponujata iste informacije: odvečnost obsega več kot 40% negotovosti razreda. Če to odštejemo od posamičnih informacijskih prispevkov, pridemo do realne ocene, ki je nekaj več kot 95%. To pomeni, da že le s tema dvema atributoma dosežemo 95% točnost pri napovedovanju razreda. Pomemben primer soodvisnosti nastopa med atributoma vonja ter oblike obročka (*ring-type*). Oblika obročka nam sama zase res odpravi kakih 30% negotovosti glede razreda, a nam ne pove ničesar, česar nam ne bi povedal že vonj gobe. Ta atribut lahko zato mirno odstranimo, če le ni v sodejavnosti s kakim drugim atributom.
Po drugi strani pa obstaja močna sodejavnost med *stalk-root* in *stalk-shape*, ki smo jo opazili že v prejšnjem razdelku. Oblika kocena nam sama po sebi ne pove praktično ničesar, v povezavi z obliko dna kocena pa odpravi več kot 55% negotovosti glede razreda. Seveda pa moramo atributa obravnavati skupaj in ne ločeno.

3.3 Interakcijski graf

Da bi dobili dober pregled nad interakcijami med atributi in razredom, lahko pregled interakcij uporabimo kot osnovo za graf, katerega vozlišča predstavljajo atribute, povezave pa interakcije. Količine, ki so v pregledu interakcij prikazane grafično, so v interakcijskem grafu prikazane numerično, kar omogoči večjo kompaktnost diagrama.
Da bi omejili kompleksnost interakcijskega grafa, ponavadi omejimo število vrisanih povezav. Graf v sliki 3 je nastal na podlagi osmih najmočnejših sodejavnosti in osmih najmočnejših soodvisnosti.



Slika 3: *Interakcijski graf prikazuje najpomembnejše interakcije v domeni kot povezave med atributi. Sodejavnosti so označene z puščičastimi povezavami, soodvisnosti pa s črtkanimi. Pod vsakim atributom je zapisan njegov informacijski prispevek.*

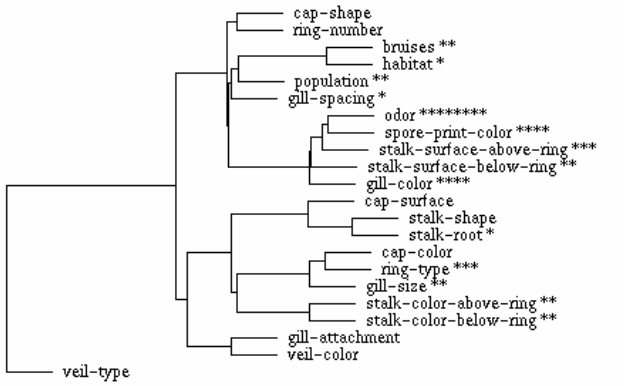
3.4 Interakcijski dendrogram

Interakcijski prispevek je pozitiven ali negativen in to odraža njegov tip, njegova absolutna vrednost pa odraža njegovo *moč*. Če je interakcija močna, je zanimiva, četudi ni tako važno, kakšnega tipa je. Na podlagi moči interakcije lahko definiramo matriko različnosti $\boldsymbol{D}$, ki opisuje različnost med atributoma A in B, takole:

$$D_{A,B} \equiv \begin{cases} \left|I(A;B;C)\right|^{-1} & \text{če } \left|I(A;B;C)\right| > c \\ c^{-1} & \text{sicer} \end{cases}$$

Konstato c izberemo tako, da nemočne interakcije ne bodo v preveliki ali premajhni meri vplivale na prikaz. Samo matriko različnosti lahko na lep način prikažemo s postopki hierarhičnega razvrščanja [4] na sliki 4.
Na podlagi takega interakcijskega dendrograma se lahko osredotočimo na odnose znotraj skupin. Najizrazitejša interakcija je med atributoma *stalk-shape* in *stalk-root*, ki opisujeta obliko kocena ter dna kocena gobe. Sama po sebi nista tako informativna, kot sta skupaj, kar smo pokazali že v prejšnjih razdelkih. Po drugi strani pa je atribut *veil-type* nepovezan z ostalimi, kar ga naredi nezanimivega za primerjavo: zavedati pa se moramo, da lahko tudi ta atribut

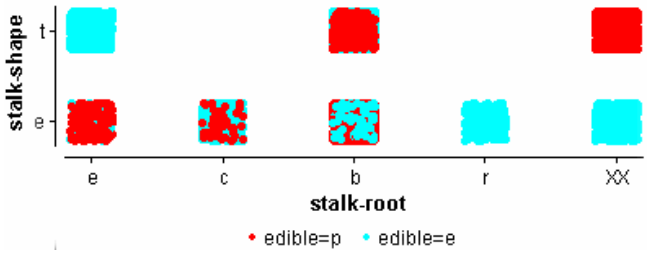
postane koristen, če bi upoštevali interakcije višjih redov. Pomen interakcijskega dendrograma je v tem, da lahko hkrati približno oriše odnose med velikim številom atributov in da izhodišče za nadaljnje raziskovanje.



Slika 4: *Razvrstitev atributov ameriških gob po moči medsebojne interakcije skupaj z razredom. Razred tu opisuje, ali je goba užitna ali ne. Število zvezdic opisuje samostojni informacijski prispevek atributa k razredu.*

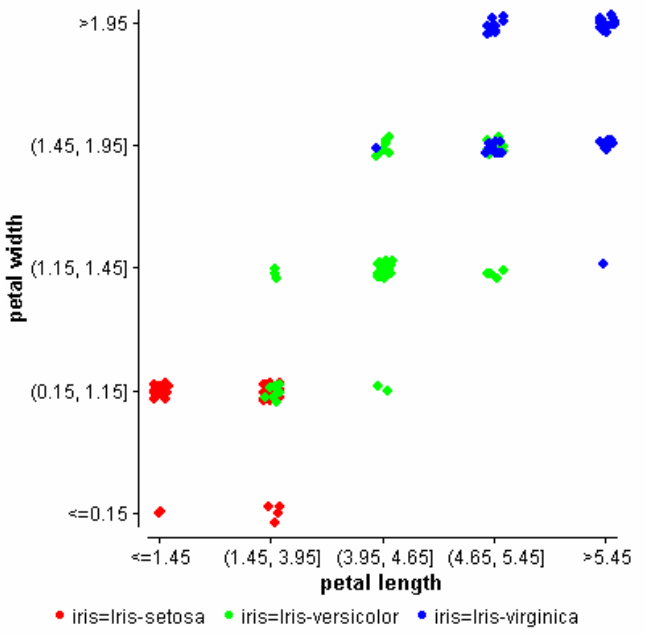
3 UPORABA INTERAKCIJ V STROJNEM UČENJU

V prejšnjem razdelku smo si ogledali postopke, ki nam na kompakten način prikažejo interakcije v podatkih. Razvili smo orodja, ki z uporabniku prijaznimi vmesniki omogočajo pregledovanje podatkov iz različnih perspektiv. Interakcijske vizualizacije lahko tu služijo za usmerjanje pogleda tja, kjer so vzorci najbolj zanimivi (slika 5).



Slika 5: *Na primeru pozitivne interakcije vidimo, da nam le hkratno obravnavanje obeh atributov v več primerih pomaga ločiti užitne od neužitnih gob.*

Interakcije nakazujejo nekatere omejitve obstoječih postopkov strojnega učenja. Z ločeno obravnavo atributov ne bomo mogli izkoristiti sinergij. Na taki obravnavi atributov temelji več popularnih metod učenja, na primer naivni Bayesov klasifikator, metoda podpornih vektorjev z linearnim jedrom ali logistična regresija, če ta ne upošteva interakcij. Po drugi strani pa nima smisla uporabljati postopkov, prilagojenih za interakcije, kot so na primer odločitvena drevesa, če interakcij ni. Postopki ocenjevanja in diskretizacije atributov so velikokrat kratkovidni, saj obravnavajo le po en atribut naenkrat. Na primeru iz slike 5 bo kratkoviden postopek ocenjevanja atributov, kot je na primer informacijski prispevek, oba atributa podcenil, saj ne bo videl sinergije. Po drugi strani pa je kratkovidni postopek diskretizacije, ki je pretvarjal zvezni atribut v diskretnega, na primeru iz slike 6 za soodvisen par atributov ustvaril veliko preveč različnih vrednosti.



Slika 6: *Na primeru vpogleda v negativno interakcijo v domeni iris vidimo, da je diskretizacija [5], ki je delovala na podlagi posamičnih atributov, ustvarila veliko več vrednosti, kot bi jih bilo treba, pri nekaterih primerih pa so združitve povzročile poslabšanje kvalitete.*

Literatura

[1] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal 27,* pp. 379-423, 623-656. 1948.

[2] W. J. McGill. Multivariate information transmission. *Psychometrika 19(2),* pp. 97-116. 1954.

[3] A. Jakulin. Interakcije med atributi v strojnem učenju. *Magistrsko delo*. Fakulteta za računalništvo in informatiko. Univerza v Ljubljani. 2003.

[4] A. Struyf, M. Hubert, P. J. Rousseeuw. Integrating Robust Clustering Techniques in S-PLUS. *Computational Statistics and Data Analysis 26,* pp. 17-37. 1997.

[5] U. M. Fayyad, K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. *Proc. of the 13th IJCAI,* pp. 1022-1027. 1993.

[6] I. Kononenko. *Strojno učenje.* Založba FE in FRI, Ljubljana. 1997.