UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Sašo Moškon

Nomogramsko iskanje podprostorov neodvisnih atributov

DIPLOMSKO DELO
NA INTERDISCIPLINARNEM UNIVERZITETNEM ŠTUDIJU

Mentor: akad. prof. dr. Ivan Bratko

Ljubljana, 2009

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00011/2009
Datum: 01.09.2009

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko ter Fakulteta za matematiko in fiziko izdaja naslednjo nalogo:

Kandidat: **SAŠO MOŠKON**

Naslov: **NOMOGRAMSKO ISKANJE PODPROSTOROV NEODVISNIH ATRIBUTOV**

NOMOGRAM-BASED SEARCH FOR SUBSPACES OF INDEPENDENT ATTRIBUTES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

V strojnem učenju je neodvisnost atributov nasplošno zaželena lastnost, ki v praksi navadno ni izpolnjena. Možno pa je, da neodvisnost velja vsaj v podprostorih celotnega učnega prostora. Tema te diplomske naloge je preučiti možnosti iskanja takih podprostorov z uporabo diskretnega odvajanja ter vizualizacije odvisnosti med diskretnimi spremenljivkami z nomogrami. Z diskretnim odvajanjem lahko iščemo podprostore, v katerih veljajo določene kvalitativne odvisnosti, te pa je mogoče uporabiti tudi pri klasifikaciji novih primerov. V okviru diplomskega dela naj bo izdelano in eksperimentalno ovrednoteno programsko orodje, ki realizira navedene funkcije.

Mentor:

akad. prof. dr. Ivan Bratko

Dekan Fakulteta za računalništvo in informatiko:

prof. dr. Franc Solina

Dekan Fakultete za matematiko in fiziko:

akad. prof. dr. Franc Forstnerič

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Sašo Moškon,

z vpisno številko 63050199,

sem avtor diplomskega dela z naslovom:

Nomogramsko iskanje podprostorov neodvisnih atributov

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom
  akad. prof. dr. Ivan Bratko
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 21.9.2009 Podpis avtorja:

Zahvala

Zahvalil bi se svojemu mentorju akademskemu profesorju Ivanu Bratku za vse nasvete ob izdelavi tega dela. Predvsem pa gre zahvala Juretu Žabkarju iz laboratorija za umetno inteligenco, ki mi je veliko svetoval in me ves čas spodbujal. Zahvalil bi se tudi punci, družini in prijateljem za potrpljenje in oporo.

Kazalo

Seznam uporabljenih kratic in simbolov

kNN - metoda najbližjih sosedov (k-Nearest Neighbours)

SVM - metoda podpornih vektorjev (Support Vector Machine)

logOR - logaritem relativnega tveganja (log Odds Ratio)

Povzetek

V diplomskem delu predstavimo *izbirne nomograme*, izboljšavo nomogramov naivnega Bayesovega klasifikatorja, ki omogočajo interaktivno raziskovanje domenskega prostora in odkrivanje pogojnih (ne)odvisnosti med atributi. Predlagamo tudi metodo iskanja okolice primera s pomočjo izbirnih nomogramov.

Najprej predstavimo izbirne nomograme, definiramo pogojne odvisnosti med atributi in podamo teoretično osnovo za odkrivanje (ne)odvisnosti med atributi z uporabo izbirnih nomogramov. Delovanje izbirnih nomogramov predstavimo na primerih in jih eksperimentalno ovrednotimo. Nato predstavimo še idejo uporabe izbirnih nomogramov pri iskanju okolice danega primera. Podamo način implementacije iskanja okolice in opišemo, kako smo te okolice uporabili pri implementaciji klasifikatorja. Klasifikator eksperimentalno ovrednotimo in opišemo razloge za slabo delovanje le-tega. V zaključku predstavimo zadane smernice za nadaljnje delo.

Ključne besede:

nomogrami, pogojne odvisnosti, okolice primera, vizualizacija

Abstract

In thesis we introduce *selective nomograms*, an improvement of nomograms for visualization of naive Bayesian classifier. Selective nomograms allow us to interactively explore the domain and discover conditional dependencies between the attributes. We also propose a classification algorithm based on the idea of selectable nomograms.

First, we introduce selective nomograms, define conditional dependencies and describe the theoretical background for discovering conditional dependencies between the attributes using selective nomograms. We present experiments and empirically evaluate selective nomograms. Then we propose the idea of using selective nomograms for searching the neighborhood of given example. We present an implementation of searching the neighborhood and describe how it can be used as a classification method. We empirically evaluate the classifier and discuss the results. Finally we propose some ideas for future work.

Key words:

nomograms, conditional dependencies, example neighborhood, visualization

Poglavje 1

Uvod

Namen tega diplomskega dela je razvoj postopka za nomogramsko iskanje podprostorov neodvisnih atributov. V okviru izdelave tega dela smo razvili izbirne nomograme in pokazali, kako lahko z njimi odkrivamo pogojne odvisnosti med atributi ter podprostore neodvisnih atributov. Izbirne nomograme smo implementirali v okolju Orange [1] in jih preizkusili. Postopek iskanja pogojnih odvisnosti smo tudi teoretično utemeljili. Idejo o izbirnih nomogramih smo uporabili pri implementaciji širjenja okolice primera. Z uporabo tega širjenja okolice smo nato razvili klasifikator in preizkusili njegovo delovanje.

Naivni Bayesov klasifikator je hiter, enostaven in pogosto uporabljan algoritem v strojnem učenju. K njegovi razširjenosti je pripomogla tudi enostavna razlaga njegovih odločitev. Za vizualizacijo modela se uporablja nomograme naivnega Bayesovega klasifikatorja [10]. Nomogram naivnega Bayesovega klasifikatorja nam za vsak atribut nariše po eno os. Na oseh so razporejene vrednosti atributov. Pozicija posamezne vrednosti prikazuje vpliv le-te na verjetnost ciljnega razreda. Ker naivni Bayesov klasifikator predpostavlja pogojno neodvisnost atributov, se nomogramska os za vsak atribut računa neodvisno od ostalih atributov. Vsak atribut je torej obravnavan individualno. Primer lahko grafično klasificiramo tako, da se na vsaki od nomogramskih osi pomaknemo na vrednost, ki ustreza primeru. Iz vsote vplivov vrednosti primera se nato izračuna verjetnost ciljnega razreda za dani primer.

Ob pomikanju po neki nomogramski osi se spreminja le vsota vplivov vrednosti, ostale nomogramske osi pa ostajajo nespremenjene. Če velja predpostavka o pogojni neodvisnosti atributov je to seveda pravilno. V primeru, da imamo opraviti s pogojno odvisnimi atributi pa temu ni tako. Recimo, da imamo v domeni pogojno odvisna atributa A in B. Ko se pomikamo po nomogramski osi atributa A, se vpliv atributa B na ciljni razred spreminja.

Analogno se tudi vpliv atributa A na ciljni razred spreminja ob pomikanju po nomogramski osi atributa B. Če se vplivi spreminjajo, bi se morale spreminjati tudi nomogramske osi. V navadnih nomogramih pa le-te ves čas ostajajo enake, saj temeljijo le na apriornih pogojnih verjetnostih. Navaden nomogram naivnega Bayesovega klasifikatorja nam torej v takem primeru ne prikaže pravih vplivov atributov na ciljni razred.

V tem diplomskem delu predstavimo izbirne nomograme [11] , ki pravilno prikažejo vplive pogojno odvisnih atributov na razred. Izbirni nomogrami so za razliko od navadnih nomogramov dinamični. Ob izbiri vrednosti nekega atributa nam omogočajo izris novega nomograma, ki upošteva to izbiro. Z izbiranjem vrednosti atributov in primerjanjem nomogramov lahko tako podrobneje raziščemo domenski prostor in celo odkrijemo pogojne odvisnosti med atributi. Tako kot nomogrami naivnega Bayesovega klasifikatorja, so tudi izbirni nomogrami preprosti in lahko razumljivi, od uporabnika pa ne zahtevajo poznavanja metod strojnega učenja. Predvsem uporabni so za eksperte z domenskim znanjem, saj jim omogočajo podrobnejšo analizo vpliva atributov na ciljni razred.

Pogojne odvisnosti med atributi so sicer pogosto obravnavana tema v strojnem učenju. Jakulin in Bratko [3] predstavita mero za zaznavanje interakcije med atributi in metodo za grafičen prikaz interakcij v domeni. S pogojnimi odvisnostmi so tesno povezane tudi Bayesove verjetnostne mreže. Struktura Bayesove mreže namreč neposredno modelira pogojne odvisnosti med atributi. Učenje strukture Bayesove mreže je torej odkrivanje pogojnih odvisnosti med atributi. Med Bayesovimi verjetnostnimi mrežami je za klasifikacijo posebej zanimiv drevesno razširjeni naivni Bayes (angl. tree augmented naive Bayes) [2]. Zanimiv je tudi pristop v [4], kjer se avtorja osredotočita na učenje neusmerjenih modelov (Markovskih mrež). Naša metoda se razlikuje v tem, da vizualizacija temelji na nomogramih in uporabniku omogoča interaktivno raziskovanje prostora in odkrivanje pogojnih (ne)odvisnosti med atributi.

V drugem poglavju so podrobneje predstavljeni naivni Bayesov klasifikator, kNN in nomogrami naivnega Bayesovega klasifikatorja. To poglavje služi predvsem za uvajanje bralca v področje in predstavitev osnovnih algoritmov, katerih poznavanje je potrebno za lažje razumevanje diplomskega dela. V tretjem poglavju so predstavljeni izbirni nomogrami ter teoretično ozadje odkrivanja (ne)odvisnosti med atributi s pomočjo izbirnih nomogramov. V četrtem poglavju prikažemo delovanje izbirnih nomogramov na nekaj domenah ter jih eksperimetalno ovrednotimo. V petem poglavju predstavimo idejo o iskanju okolice primera s pomočjo izbirnih nomogramov in predstavimo klasifikacijo, ki temelji omenjenemu širjenju okolice. Klasifikator tudi empirično ovrednotimo.

Ker se klasifikator slabo izkaže ob koncu poglavja podamo razloge za težave le-tega.

Poglavje 2

Pregled področja

Diplomsko delo spada na področje umetne inteligence. Ožje spada na področje strojnega učenja. Za razumevanje tega diplomskega dela je potrebno podrobneje poznati delovanje naivnega Bayesovega klasifikatorja, nomogramov in algoritma kNN.

2.1 Naivni Bayesov klasifikator

Bayesov klasifikator [6] izračuna pogojne verjetnosti za vsak razred pri danih vrednostih atributov za dani novi primer, ki ga želimo klasificirati. Bayesov klasifikator, ki natančno izračuna pogojne verjetnosti razredov je optimalen, v tem smislu, da minimizira pričakovano napako. Ker natančno računanje vseh pogojnih verjetnosti ni vedno mogoče in bi bilo pogosto prezahtevno, se uporabljajo določene predpostavke. Naivni Bayesov klasifikator predpostavi pogojno neodvisnost atributov pri danem razredu. To omogoči, da učna množica običajno zadošča za zanesljivo oceno vseh potrebnih verjetnosti za izračun končne pogojne verjetnosti vsakega razreda. Uporabili bomo naslednje oznake: $P(r_k)$ apriorna verjetnost razreda r_k
$V =< v_1, \cdots, v_a >$ vektor vrednosti za vse atribute
$P(r_k|V)$ verjetnost razreda pri danih vrednostih atributov
$P(r_k|v_i)$ verjetnost razreda pri dani vrednosti i-tega atributa
Verjetnosti razredov se računa po enačbi

$$P(r_k|V) = P(r_k) \prod_{i=1}^{a} \frac{P(r_k|v_i)}{P(r_k)} \tag{2.1}$$

Naloga učnega algoritma je s pomočjo učne množice podatkov aproksimirati verjetnosti na desni strani enačbe. Znanje naivnega Bayesovega klasifikatorja

je torej tabela aproksimacij apriornih verjetnosti razredov $P(r_k), k = 1, \cdots, n_0$ in tabela pogojnih verjetnosti razredov $r_k, k = 1, \cdots, v_0$ pri dani vrednosti v_i atributa $A_i, i = 1, \cdots, a : P(r_k|v_i)$.

Za ocenjevanje apriornih verjetnosti se pogosto uporablja Laplaceov zakon zaporednosti

$$P(r_k) = \frac{N_k + 1}{N + n_0} \tag{2.2}$$

Za ocenjevanje pogojnih verjetnosti se pogosto uporablja m-ocena

$$P(r_k|v_i) = \frac{N_{k,i} + mP(r_k)}{N_i + m}, \tag{2.3}$$

kjer je $N_{k,i}$ število učnih primerov iz razreda r_k in z vrednostjo i-tega atributa v_i ter N_i število vseh učnih primerov z vrednostjo i-tega atributa v_i.

2.2 Nomogrami

Nomograme je leta 1891 prvič predstavil francoski matematik Maurice d'Ocagne. Njihov namen je bil omogočiti uporabniku, da grafično izračuna rezultat enačbe, ne da bi pri tem moral kaj preračunavati. V strojnem učenju so nomograme prvič uporabili Lubsen in soavtorji [9] pri vizualizaciji modelov logistične regresije. Pokazali so uporabnost nomogramov na medicinski domeni. Njihov nomogram je bil načrtovan tako, da ga je bilo mogoče natisniti na list papirja in uporabljati brez računalnika. Poznamo tudi nomograme za vizualizacijo metode podpornih vektrojev. Vizualizacija z nomogrami za to metodo je predstavljena v [5]. Nas najbolj zanimajo nomogrami naivnega Bayesovega klasifikatorja [10].

Za izris nomogramov naivnega Bayesovega klasifikatorja moramo najprej izračunati logaritem relativnega tveganja (log odds ratio)

$$logOR(A_i) = log\frac{P(A_i|r_k)}{P(A_i|\bar{r_k})} = log\frac{\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)}}{\frac{P(r_k)}{P(\bar{r_k})}} \tag{2.4}$$

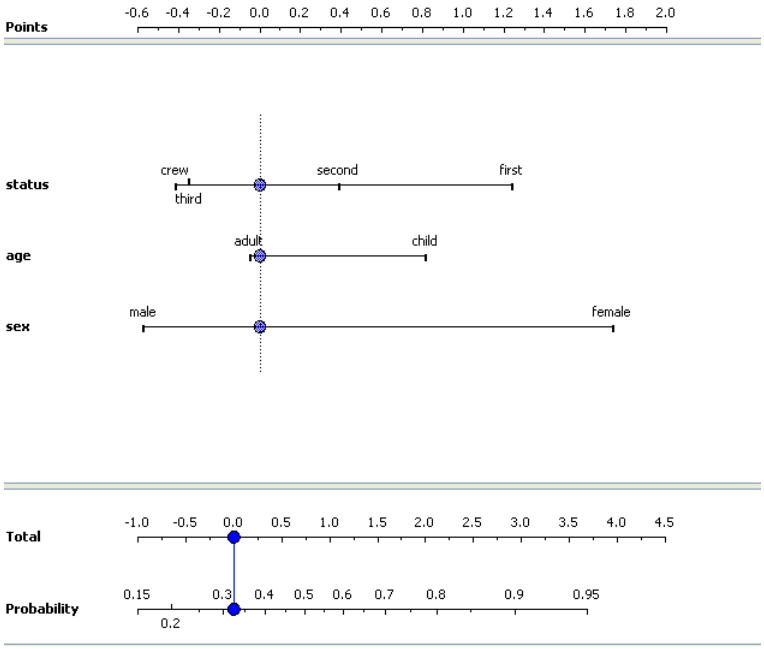
Tako ocenimo kako posamezna vrednost atributa vpliva na verjetnost ciljnega razreda. Pozitiven logOR pomeni, da je verjetnost ciljenga razreda pri dani vrednosti atributa q_i večja kot apriorna verjetnost ciljnega razreda. Negativen pa, da je le-ta manjša.

Iz tako dobljenih logOR vrednosti lahko nato izračunamo verjetnost ciljnega razreda. Najprej seštejemo prispevke posameznih atributov

$$F(r_k|X) = \sum logOR(A_i) \tag{2.5}$$

nato izračunamo verjetnost ciljnega razreda z naslednjo formulo

$$P(r_k|X) = [1 + e^{-logP(r_k)/(1-P(r_k))-F(r_k|X)}]^{-1} \tag{2.6}$$



Slika 2.1: Primer nomograma na podatkih o preživetju nesreče titanica

2.3 Algoritem kNN

Algoritem kNN [7] je klasifikator, ki klasificira na podlagi najbližjih sosedov. Najpreprostejša varianta algoritma najbližjih sosedov kot znanje uporablja kar množico vseh učnih primerov. Učni algoritem si torej le zapomni vse primere. Ker učenja skorajda ni, pravimo tej vrsti učenja tudi leno učenje. Pri klasifikaciji novega primera se iz učne množice poišče k najbolj podobnih (najbližjih) primerov. Nov primer klasificiramo v razred, ki mu pripada največ

bližnjih sosedov. Pri tem je potrebno zaradi ustrezne metrike v prostoru atributov normalizirati vrednosti zveznih atributov in definirati razdaljo med vrednostmi vsakega diskretnega atributa. Pri klasifikaciji se navadno uporablja tudi uteževanje primerov glede na njihovo oddaljenost od primera, ki ga klasificiramo. Uporablja se različne razdalje: za zvezne atribute najpogosteje evklidsko in za diskretne Hammingovo razdaljo.

Hammingova razdalja

Hammingova razdalja nam pove, na koliko mestih se vektorja razlikujeta. Za binarna vektorja $x = (x_1, ...x_n)$ in $y = (y_1, ...y_n)$ enakih dolžin je definirana kot

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.7}$$

Če si posamezen primer v strojnem učenju predstavljamo kot vektor vrednosti $(v_1, ...v_n)$ lahko Hammingovo razdaljo med primeri definiramo kot

$$d(x, y) = \sum_{i=1}^{n} f(x_i, y_i) \tag{2.8}$$

kjer je

$$f(x_i, y_i) = \begin{cases} 0 & if \quad x_i = y_i \\ 1 & sicer \end{cases} \tag{2.9}$$

Hammingova razdalja nam torej pove, po kolikih atributih se primera razlikujeta.

Za primer vzemimo vektorja $x = (2, 3, 1, 1)$ in $y = (2, 3, 7, 1)$, ki naj predstavljata primera v domeni s štirimi atributi. Hammingova razdalja teh dveh vektorjev je ena. Razlikujeta se na tretjem mestu v vektorju.

$$\begin{aligned} x &= (2, 3, \mathbf{1}, 1) \\ y &= (2, 3, \mathbf{7}, 1) \end{aligned} \tag{2.10}$$

Pomembno je torej le, da se vrednosti razlikujeta, ne pa za koliko se razlikujeta. To lastnost je uporabna predvsem v diskretnih domenah, kjer imamo pogosto opravka z nominalnimi vrednostmi atributov in razdalje med posameznimi vrednostmi ne poznamo.

Poglavje 3

Izbirni nomogrami

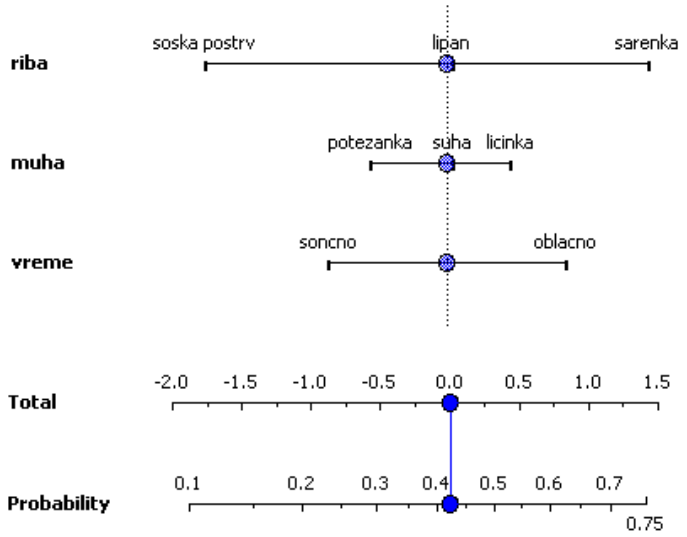
3.1 Uvod v izbirne nomograme

Nomogrami naivnega Bayesovega klasifikatorja [10] so zelo koristni za vizualizacijo odločitev naivnega Bayesovega klasifikatorja. Omogočajo nam pregled nad tem kako posamezne vrednosti atributov vplivajo na verjetnost ciljnega razreda. Vendar se v podatkih pogosto skriva več informacije kot jo lahko prikaže nomogram naivnega Bayesovega klasifikatorja. Na sliki 3.1 vidimo nomogram naivnega Bayesovega klasifikatorja za podatke o ulovih pri muharjenju.

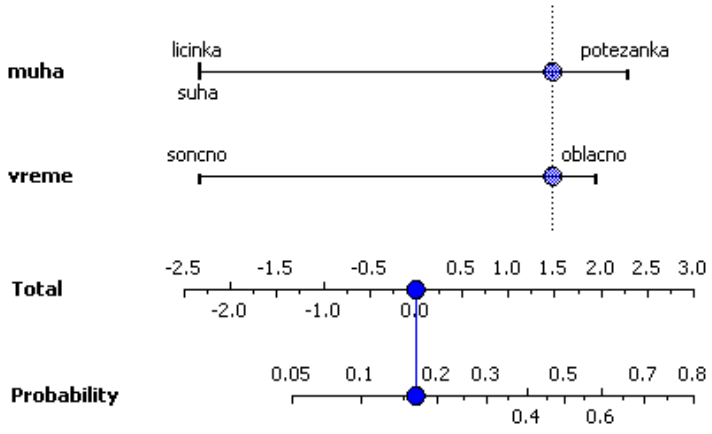
Domeno sestavljajo atribut *riba* z vrednostmi <*soška postrv, lipan, šarenka*>, atribut *muha* z vrednostmi <*potezanka, suha, ličinka*> ter atribut *vreme* z vrednostima <*sončno, oblačno*>. Vsi atributi so diskretni. Razred je diskreten atribut *prijem* z vrednostima <*da,ne*>.

Iz nomograma lahko razberemo, da je najbolje loviti v oblačnem vremenu, najboljša vaba je imitacija ličinke, najpogostejši ulov pa šarenka. Imamo torej vse potrebne informacije za uspešen lov, le še pogledamo vremensko napoved, pripravimo opremo in se odpravimo za vodo. Prvi dan se vrnemo z nekaj šarenkami, drugi dan s še več šarenkami tudi tretji dan se kljub sončnemu vremenu odpravimo na lov in uplenimo nekaj šarenk in lipana. Dnevi se vrstijo in naveličani smo uvoženih šarenk, želimo si prijema domače, soške postrvi. A ta noče prijeti. Iz nomograma smo razbrali, da je verjetnost prijema soške postrvi majhna a ne nična. Morda smo kaj spregledali? Naveličani šarenk se ponovno posvetimo nomogramom. Tokrat si poglejmo podatke le za soško postrv. Nomogram je prikazan na sliki 3.2.

Zanimivo, ta nomogram nam pove, da ličinka ni primerna vaba za soško postrv in da je ob sončnem vremenu bolje doma pokositi travo kot izgubljati



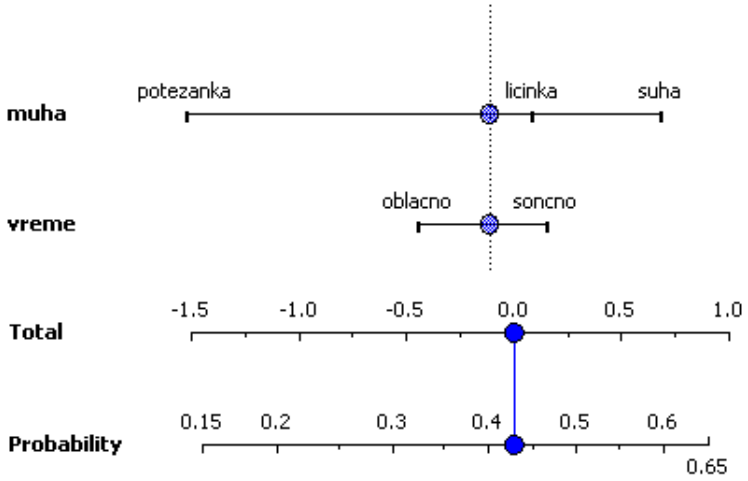
Slika 3.1: Primer nomograma naivnega Bayesovega klasifikatorja za podatke o ulovih pri muharjenju



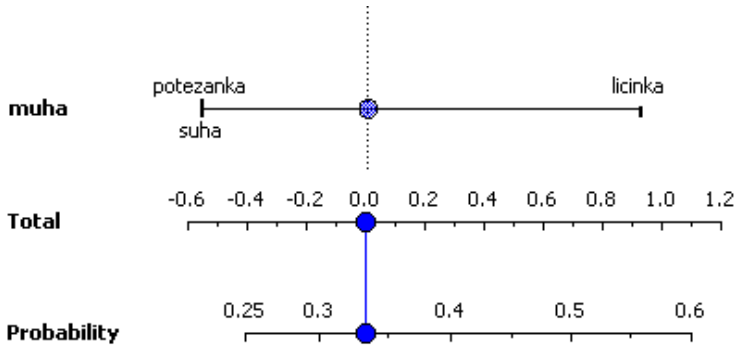
Slika 3.2: Primer nomograma naivnega Bayesovega klasifikatorja za podatke o ulovih pri muharjenju, omejene le na ulove soške postrvi

čas za vodo. Na lov se odpravimo tik pred nevihto, na predvrvico navežemo potezanko in kmalu se vrnemo domov z nasmeškom na obrazu. Začetni nomogram nas je očitno zavedel, ličinka, ki naj bi bila najbolj uspešna vaba, se je izkazala za zelo slabo pri lovu soške postrvi. Začetni nomogram ni bil napačen. Pokazal je tisto kar lahko naivni Bayes izračuna na danih podatkih. Vendar

naivni Bayes predpostavlja pogojno neodvisnost atributov, ribam pa ni vseeno katero vabo jim ponudimo. Če si pogledamo nomogram na sliki 3.3 vidimo, da jim tudi za vreme ni vseeno. Lipan namreč najraje prijema na suho muho v sončnih dneh. Če se ga že lotimo v oblačnem vremenu pa je za vabo vendarle bolje uporabiti ličinko, saj na suho muho ne bo prijel, kot je razvidno iz nomograma na sliki 3.4.



Slika 3.3: Primer nomograma naivnega Bayesovega klasifikatorja za podatke o ulovih pri muharjenju, omejene le na ulove lipana



Slika 3.4: Primer nomograma naivnega Bayesovega klasifikatorja za podatke o ulovih pri muharjenju, omejene le na ulove lipana v oblačnem vremenu

Podatki lahko torej skrivajo veliko več informacij, kot jih lahko nomo-

gram prikaže. Hkrati to pomeni, da je v podatkih lahko veliko več relacij, kot jih naivni Bayes odkrije. Naivni Bayesov klasifikator bi se na podatkih iz prejšnjega primera sicer solidno izkazal, a očitno je, da se da tudi bolje. Zmotijo ga seveda pogojne odvisnosti med atributi.

3.2 Pogojne odvisnosti med atributi

Ko velja predpostavka o pogojni neodvisnosti med atributi, je naivni Bayesov klasifikator optimalen. V tem primeru nam nomogram naivnega Bayesovega klasifikatorja nudi popolno informacijo o vplivu atributov na razred. Poigravanje s podatki, kakršno smo prikazali v prejšnjem razdelku, ne privede do novih spoznanj. Predpostavka o pogojni neodvisnosti v realnih podatkih le redko drži.

Če poznamo izid c sta dogodka a in b pogojno neodvisna ko velja

$$P(a, b|c) = P(a|c) \times P(b|c) \tag{3.1}$$

Primer pogojne odvisnosti med atributoma je podan v tabeli 3.1. Čeprav sta vrednosti atributov neodvisni, postaneta močno pogojno odvisni ko je podana vrednost razreda. Na sliki 3.5 je nomogram naivnega Bayesovega klasifikatorja za podatke iz tabele 3.1. Če sklepamo po tem nomogramu, je vrednost

X1	**X2**	**Y**
0	0	0
1	0	1
0	1	1
1	1	0

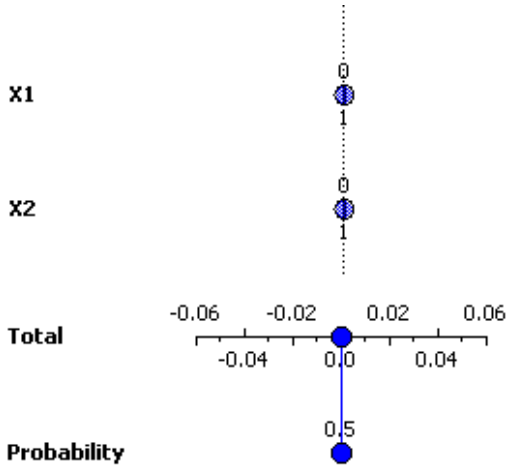
Tabela 3.1: Primer pogojne odvisnosti med atributoma X1 in X2

razreda popolnoma neodvisna od vrednosti posameznega atributa. Vendar to ni res. Kot lahko razberemo iz tabele 3.1 velja med atributoma in razredom naslednja odvisnost:

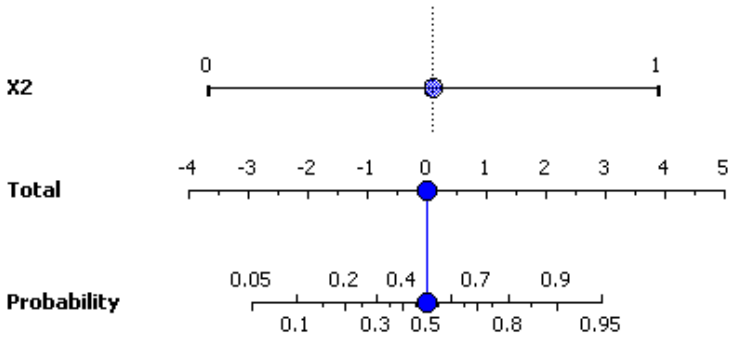
$$Y = (X1 \neq X2) \tag{3.2}$$

Po zgledu iz poglavja 3.1, bi moral torej nomogram za nek podprostor podatkov razkriti kaj več. Izberimo le podatke, ki imajo vrednost atributa X1 enako 1. Nomogram naivnega Bayesovega klasifikatorja za take podatke je

prikazan na sliki 3.6. Opazimo, da se je vpliv atributa *X2* na razred močno spremenil.



Slika 3.5: Primer nomograma naivnega Bayesovega klasifikatorja za podatke iz tabele 3.1



Slika 3.6: Primer nomograma naivnega Bayesovega klasifikatorja za primere iz tabele 3.1 ki imajo vrednost atributa X1 enako 1

Atributa sta torej pogojno odvisna če poznavanje vrednosti enega atributa spremeni vpliv vrednosti drugega atributa na razred. To pomeni, da se spremeni nomogramska os za ta atribut. Spremenijo se lahko le razdalje med vrednostmi atributov na nomogramski osi ali celo vrstni red le-teh. Zanimivi so predvsem dogodki, ko se spremeni vrstni red vrednosti atributov na nomogramskih oseh. To se je dogajalo v primeru, ki smo ga uporabili v razdelku

3.1, ko smo ugotovili, da se za različne ribe dobro obnesejo različne vabe. Upoštevati je potrebno tudi dejansko velikost spremembe. Ni namreč vseeno ali se na izbirnem nomogramu ob izbiri neke vrednosti atributa na obravnavani osi zamenjata vrednosti, ki sta bili v originalnem nomogramu zelo skupaj, ali vrednosti, ki sta bili precej oddaljeni. Pomembno je tudi za koliko se spremeni vpliv posamezne vrednosti na končni razred. Ko govorimo o razdalji mislimo na razliko v logaritmu relativnega tveganja. Pomembno je tudi upoštevati zastopanost podprostora, za katerega smo izrisali nomogram naivnega Bayesovega klasifikatorja. Če dobimo zelo različne nomogramske osi za nek atribut, pa so te izračunane le na podlagi malega števila primerov, iz tega seveda ne moremo sklepati na pogojne odvisnosti med atributi.

3.3 Teoretično ozadje odkrivanja podprostorov neodvisnih atributov

Vrednosti logOR za i-to vrednost atributa A izračunamo po enačbi

$$logOR(A_i) = log\frac{P(A_i|r_k)}{P(A_i|\bar{r_k})} = log\frac{\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)}}{\frac{P(r_k)}{P(\bar{r_k})}} \tag{3.3}$$

Če izberemo j-to vrednost atributa B izračunamo logOR po sledeči formuli

$$logOR(A_i|B_j) = log\frac{P(A_i, B_j|r_k)}{P(A_i, B_j|\bar{r_k})} = log\frac{\frac{P(r_k|A_i,B_j)}{P(\bar{r_k}|A_i,B_j)}}{\frac{P(r_k|B_j)}{P(\bar{r_k}|B_j)}} \tag{3.4}$$

Če velja predpostavka o pogojni neodvisnosti atributov lahko enačbo 3.4 poenostavimo. Zaradi pogojne neodvisnosti namreč velja

$$P(r_k|A_i, B_j) = P(r_k|A_i) \tag{3.5}$$

iz tega sledi

$$logOR(A_i|B_j) = log\frac{P(A_i, B_j|r_k)}{P(A_i, B_j|\bar{r_k})} = log\frac{\frac{P(r_k|A_i,B_j)}{P(\bar{r_k}|A_i,B_j)}}{\frac{P(r_k|B_j)}{P(\bar{r_k}|B_j)}} = log\frac{\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)}}{\frac{P(r_k|B_j)}{P(\bar{r_k}|B_j)}} \tag{3.6}$$

Enačbi za izračun vrednosti $logOR(A_i)$ in $logOR(A_i|B_j)$ se torej razlikujeta le v imenovalcu. Števec je pri obeh enak. To pomeni, da se vrstni red vrednosti na nomogramski osi, ko velja predpostavka o pogojni neodvisnosti atributov,

ne more spremeniti. To sledi direktno iz lastnosti logaritma. Ker delamo z verjetnostmi, ki so seveda vedno pozitivne, vedno velja, da večje število v števcu pomeni večjo vrednost logaritma. To velja ne glede na vrednost v imenovalcu, ki je, kot smo že poudarili, vedno pozitivna. Zaradi te monotone relacije med vrednostmi števca in vrednostmi logOR imenovalec ne more vplivati na vrstni red na nomogramski osi. Torej lahko trdimo, da izbira vrednosti enega atributa, ob veljavnosti predpostavke o neodvisnosti med atributi, ne more spremeniti vrstnega reda vrednosti na nomogramski osi kateregakoli drugega atributa. Iz tega sledi, da lahko iz spremebe vrstnega reda vrednosti na nomogramski osi nekega atributa *A1*, po tem ko smo izbrali vrednost atributa *A2*, sklepamo na odvisnost med atributoma. Sedaj si poglejmo še razdaljo na nomogramski osi. Osredotočimo se na razdaljo med dvema vrednostima atributa A, vrednostima A_i in A_j. Njuno razdaljo na nomogramski osi v osnovnem nomogramu naivnega Bayesovega klasifikatorja lahko izračunamo kot

$$\begin{aligned} d &= logOR(A_i) - logOR(A_j) \\ &= log\frac{\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)}}{\frac{P(r_k)}{P(\bar{r_k})}} - log\frac{\frac{P(r_k|A_j)}{P(\bar{r_k}|A_j)}}{\frac{P(r_k)}{P(\bar{r_k})}} \\ &= log\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)} - log\frac{P(r_k)}{P(\bar{r_k})} - log\frac{P(r_k|A_j)}{P(\bar{r_k}|A_j)} + log\frac{P(r_k)}{P(\bar{r_k})} \\ &= log\frac{P(r_k|A_i)}{P(\bar{r_k}|A_i)} - log\frac{P(r_k|A_j)}{P(\bar{r_k}|A_j)} \end{aligned}$$

Kot lahko razberemo iz zgornje izpeljave, je tudi razdalja med atributoma odvisna le od imenovalca. Že v enačbi 3.6 smo pokazali, da se ob veljavni predpostavki o neodvisnosti atributov, ob izbiri vrednosti enega izmed atributov v enačbah za izračun novih logOR spremeni le imenovalec. Iz tega lahko zaključimo, da se ob izbiri vrednosti enega izmed atributov in veljavni predpostavki o pogojni neodvisnosti atributov, razdalje na nomogramskih oseh ostalih atributov ne spremenijo. To pomeni, da lahko iz velike spremembe razdalje med vrednostmi atributa A na nomogramski osi po tem ko smo izbrali neko vrednost atributa B sklepamo na pogojno odvisnost med tema dvema atributoma.

Primer

Domena naj vsebuje binarni razred Y ter dva binarna atributa X_1 in X_2. Za primer vzemimo naslednje podatke o verjetnostih:

$$\begin{aligned} P(Y=1) &= 0.6 \\ P(Y=1|X_1=1) &= 0.5 \\ P(Y=1|X_1=0) &= 0.8 \\ P(Y=1|X_2=1) &= 0.45 \\ P(Y=1|X_2=0) &= 0.78 \end{aligned}$$

Za izračun vrednosti logOR potrebujemo še pogojne verjetnosti $P(Y \neq 1), P(Y \neq 1|X_2=1)$ in $P(Y \neq 1|X_2=0)$. Izračunamo jih iz že znanih verjetnosti

$$\begin{aligned} P(Y \neq 1)1 - P(Y=1) &= 1 - 0.6 = 0.4 \\ P(Y \neq 1|X_2=1) &= 1 - P(Y=1|X_2=1) = 1 - 0.45 = 0.55 \\ P(Y \neq 1|X_2=0) &= 1 - P(Y=1|X_2=0) = 1 - 0.78 = 0.22 \end{aligned}$$

S temi podatki lahko izračunamo logOR vrednosti za atribut X_2

$$logOR(X_2=1) = log\frac{\frac{P(Y=1|X_2=1)}{P(Y\neq1|X_2=1)}}{\frac{P(Y=1)}{P(Y\neq1)}} = log\frac{\frac{0.45}{0.55}}{\frac{0.6}{0.4}}0 = -0.606$$

$$logOR(X_2=0) = log\frac{\frac{P(Y=1|X_2=0)}{P(Y\neq1|X_2=0)}}{\frac{P(Y=1)}{P(Y\neq1)}} = log\frac{\frac{0.78}{0.22}}{\frac{0.6}{0.4}}0 = 0.860$$

Vrednost atributa $X_2 = 1$ torej poveča verjetnost razreda $Y = 1$, vrednost $X_2 = 0$ pa jo zmanjša. Sedaj izberimo vrednost atributa $X_1 = 1$. To pomeni, da od sedaj upoštevamo le primere, ki imajo vrednost atributa $X_1 = 1$. Ker sta atributa X_1 in X_2 pogojno neodvisna, ostanejo pogojne verjetnosti za atribut X_2 enake. Sicer bi bila atributa pogojno odvisna. Zaradi izbire vrednosti atributa X_1 se spremeni verjetnost razreda $Y = 1$. Po enačbi 3.6 izracunamo nove logOR vrednosti za atribut X_2

$$logOR(X_2=1|X_1=1) = log\frac{\frac{P(Y=1|X_2=1)}{P(Y\neq1|X_2=1)}}{\frac{P(Y=1|X_1=1)}{P(Y\neq1|X_1=1)}} = log\frac{\frac{0.45}{0.55}}{\frac{0.5}{0.5}}0 = -0.200$$

$$logOR(X_2=0|X1=1) = log\frac{\frac{P(Y=1|X_2=0)}{P(Y\neq1|X_2=0)}}{\frac{P(Y=1|X_1=1)}{P(Y\neq1|X_1=1)}} = log\frac{\frac{0.78}{0.22}}{\frac{0.5}{0.5}}0 = 1.266$$

Vrednosti logOR so se spremenile, vendar vrstni red vrednosti atributa ostaja enak. logOR za vrednost $X_2 = 1$ je namreč pred in po izbiri manjši od logOR za vrednost $X_2 = 0$. Izračunajmo še razdaljo med vrednostima pred izbiro $X_1 = 1$

$$logOR(X_2 = 1) - logOR(X_2 = 0) = -0.606 - 0.860 = -1.466 \tag{3.7}$$

in po izbiri $X_1 = 1$

$$logOR(X_2 = 1|X_1 = 1) - logOR(X_2 = 0|X_1 = 1) = -0.200 - 1.266 = -1.466 \tag{3.8}$$

Po pričakovanjih se vrednosti ujemata.

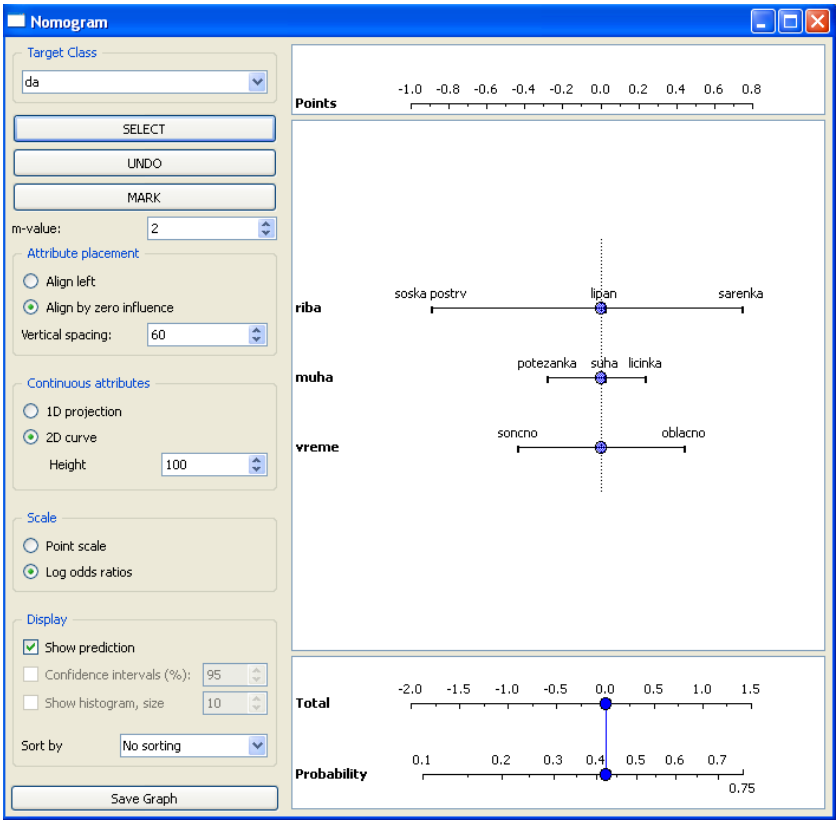
3.4 Implementacija izbirnih nomogramov

V okviru tega diplomskega dela smo implementirali izbirne nomograme naivnega Bayesovega klasifikatorja. Izbirni nomogrami so nomogrami, ki poleg funkcionalnosti navadnih nomogramov uporabniku omogočajo še določanje posameznih vrednosti atributov in s tem analizo izbranih podprostorov podatkov. Izbirne nomograme smo implementirali v programskem jeziku Python kot del orodja ORANGE [1]. Implementacija je razširitev že obstoječega čarovnika za risanje nomogramov v odprtokodnem orodju ORANGE.

Izgled okna izbirnih nomogramov je prikazan na sliki 3.7. Poleg gumbov in polj za nastavitve, ki so bili že del čarovnika, smo dodali še Gumbe *SELECT*, *UNDO* in *MARK*. Dodano je tudi polje za določanje m-ocene, ki naj se uporablja pri preračunavanju pogojnih verjetnosti. Uporaba izbirnih nomogramov je zelo enostavna. Uporabnik mora le s kurzorjem miške premakniti drsnik izbranega atributa na določeno vrednost in klikniti gumb *SELECT*. če drsnik nastavimo med dve vrednosti, se ob pritisku gumba *SELECT* izpiše opozorilo, da nismo izbrali nobene vrednosti. Gumb *UNDO* služi za razveljavljanje prejšnjih izbir. Z vsakim pritiskom tega gumba se razveljavi ena prejšnja izbira, ki smo jo naredili s pritiskom gumba *SELECT*. Gumb *MARK* služi za izbiranje intervalom pri zveznih atributih.

Funkcija gumba *SELECT*

Čarovnik si zapomni, kateri drsnik smo nazadnje premikali. Ob pritisku gumba *SELECT* se izbrana vrednost atributa ali izbrani interval doda v posebno tabelo pogojev. Nato se tabele primerov odstrani vse primere, ki ne ustrezajo pogojem v tabeli pogojev. Na teh primerih se ponovno izračuna vse potrebno



Slika 3.7: Okno čarovnika "Izbirni nomogrami"

za izris nomograma naivnega Bayesovega klasifikatorja. Izriše se nomogram, v konzolo se izpišejo pogoji iz tabele pogojev in število primerov, ki ustrezajo tem pogojem.

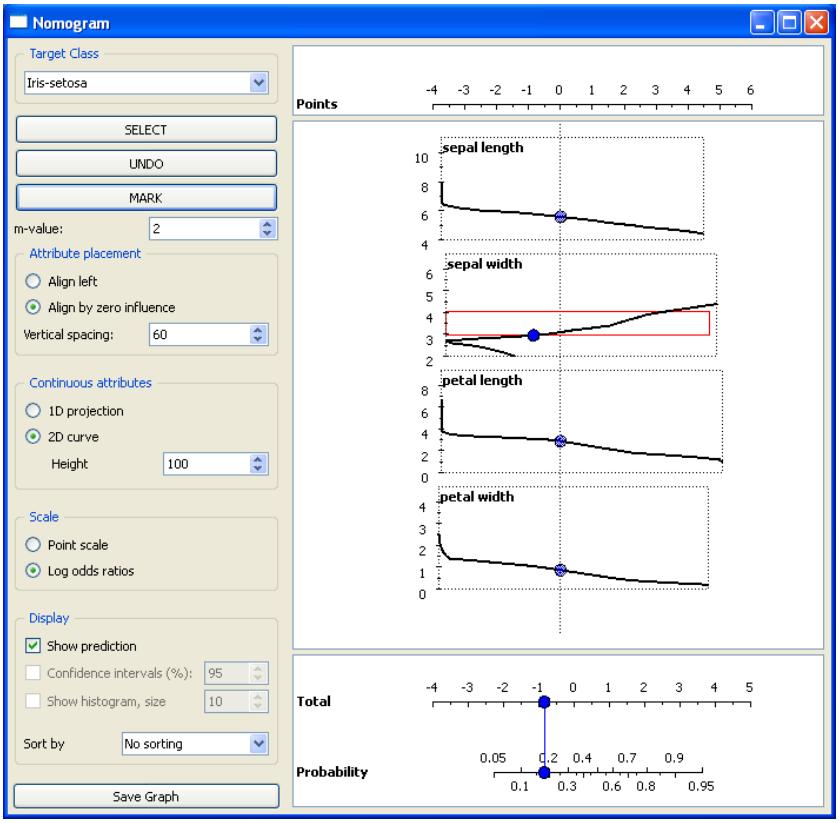
Funkcija gumba *UNDO*

Gumb *UNDO* iz tabele pogojev odtrani zapis, ki je bil v tabelo dodan zadnji. Gumb lahko uporabimo tudi večkrat. Če je ob pritisku gumba tabela pogojev prazna, nas čarovnik na to utrezno opozori.

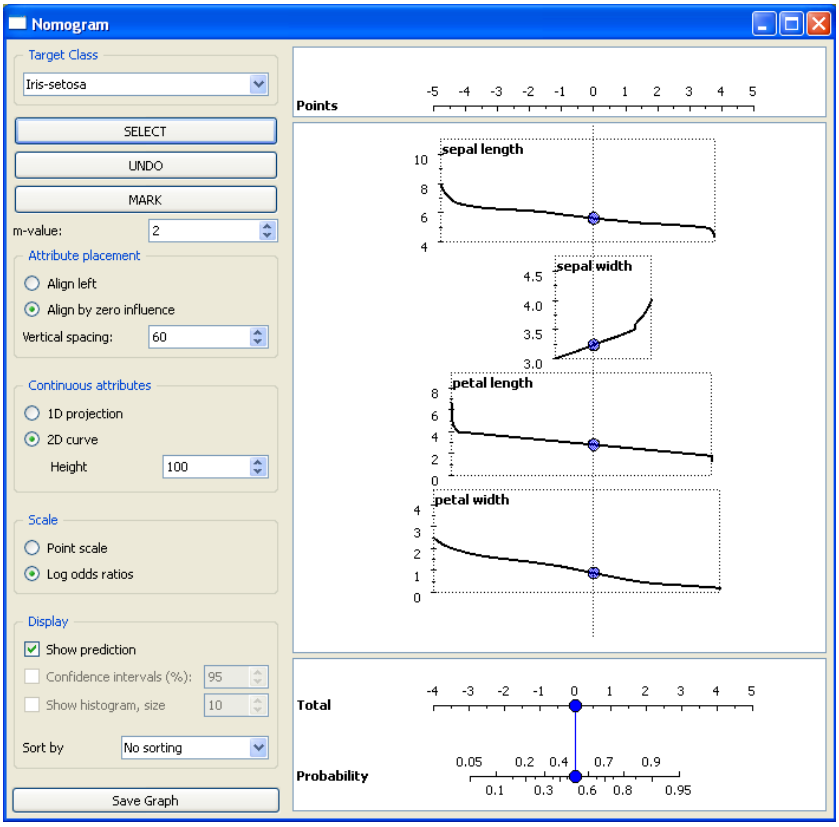
Funkcija gumba *MARK*

Gumb *MARK* služi za označevanje intervalov na 2D krivuljah zveznih atributov v nomogramu. Pri zveznih atributih so vrednosti navadno zelo raznolike in nesmiselno bi bilo izbirati posamezne vrednosti. Za izbiro intervala vred-

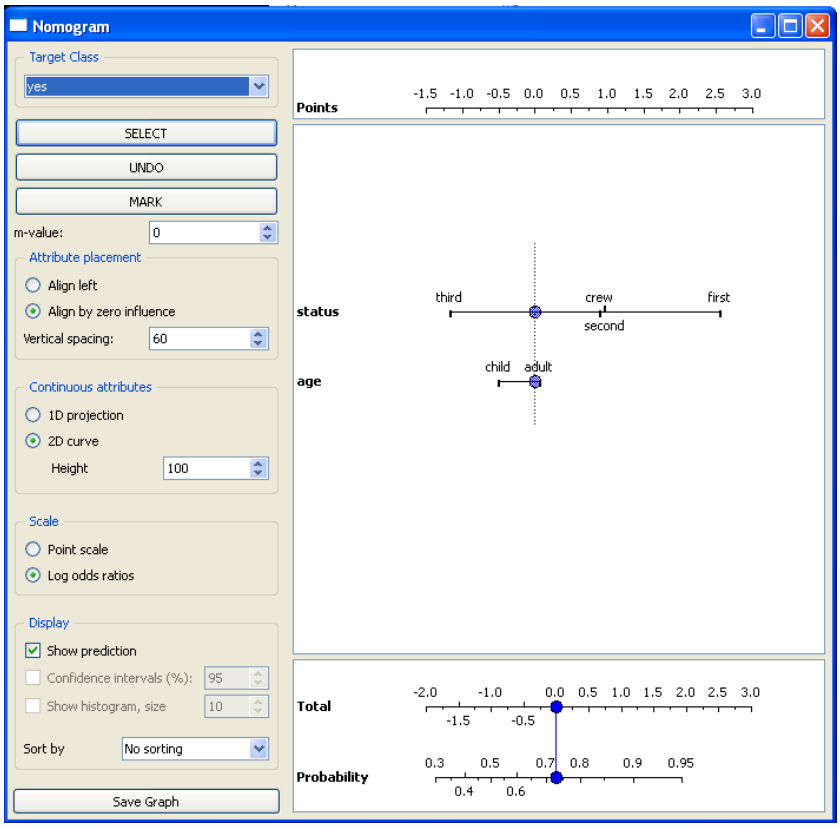
nosti zveznega atributa mora biti izbran prikaz zveznih atributov z 2D krivuljo. Gumb uporabljamo tako da drsnik najprej pomaknemo na začetek željenega intervala in kliknemo gumb. Na mesto kjer je drsnik se izriše kratka črtica. Nato drsnik pomaknemo na konec željenega intervala in ponovno kliknemo gumb. Okoli izbranega intervala se izriše kvadrat, ki označuje interval. Z uporabo gumba *SELECT* lahko nato izrišemo nomogram za podatke, ki imajo vrednost obravnavanega atributa znotraj izbranega intervala. Na sliki 3.8 je prikazan izbran interval *sepal width* $\in$ *[2.2 , 3.6]*, na sliki 3.9 pa izrisan nomogram po pritisku gumba *SELECT*.



Slika 3.8: Primer nomograma na zveznih podatkih iris z označenim intervalom *sepal width = [2.2 , 3.6]*



Slika 3.9: Primer nomograma na podatkih iris z izbranim intervalom *sepal width = [2.2 , 3.6]*



Slika 3.10: Primer nomograma na podatkih o preživetju nesreče titanica

Poglavje 4

Eksperimentalno ovrednotenje izbirnih nomogramov

Kot smo ugotovili v razdelku 3.2, lahko pogojne odvisnosti med atributi odkrijemo že z opazovanjem nomogramskkih osi medtem, ko raziskujemo prostor podatkov. Dva atributa sta pogojno neodvisna, če izbira vrednosti enega atributa ne vpliva na vrstni red in niti na razdalje na nomogramski osi drugega atributa. Seveda moramo, ko imamo opravka z realnimi podatki, te trditve vzeti nekoliko z rezervo. Majhne spremembe v razdalji še ne pomenijo odvisnosti med atributi. Tudi sprememba vrstnega reda atributov nam ne pove veliko, če ni podprta z zadostnim številom primerov.

4.1 Delovanje na domeni XOR

4.1.1 Opis domene XOR

Domena XOR je sestavljena iz dveh diskretnih atributov X1 in X2 ter diskretnega razreda Y. Zveza med atributoma in razredom je podana z enačbo

$$Y = (X1 \neq X2) \tag{4.1}$$

Gre za ekskluzivni ali. Vrednost razreda Y je torej enaka ena, ko natanko eden izmed atributov X1 in X2 zavzame vrednost ena. Atributa X1 in X2 sta med seboj močno pogojno odvisna.

4.1.2 Eksperiment na domeni XOR

Naivni Bayesov klasifikator predpostavlja pogojno neodvisnost, zato je na taki domeni neuspešen. Posledično nam tudi nomogram naivnega Bayesovega klasifikatroja za podatke iz te domene ne pove ničesar. Nomogram je prikazan na sliki 3.5. Na tem mestu se izkažejo izbirni nomogrami. Če namreč izberemo eno izmed vrednosti atributa X1, nam novi nomogram, ki temelji le na podatkih z izbrano vrednostjo atributa X1, razkrije nove informacije o atributih. Kot je razvidno iz slike 3.6, nam v tem primeru vrednost drugega atributa natančno določa vrednost razreda. Analogno velja za izbiro ene izmed vrednosti atributa X2. Ob izbiri vrednosti enega izmed atributov se torej nomogramska os drugega atributa močno spremeni. To nam pove, da sta atributa med seboj močno odvisna. Tako smo z izbirnimi nomogrami odkrili odvisnost med atributoma.

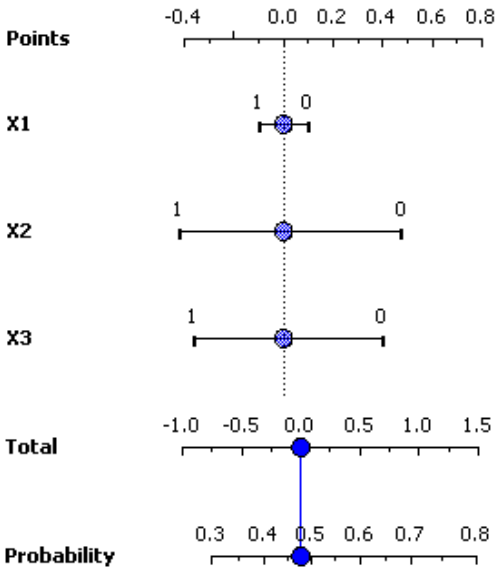
4.2 Delovanje na umetni domeni UM1

4.2.1 Opis umetne domene UM1

Za prikaz delovanja izbirnih nomogramov smo ustvarili umetno domeno s tremi diskretnimi atributi *X1, X2 in X3* ter diskretnim razredom *Y*. Vsi imajo zalogo vrednosti $0, 1$. Primere smo dodajali tako, da smo ustvarili nekatere pogojne odvisnosti med atributi. Atributa *X2* in *X3* sta pogojno neodvisna. Oba pa sta odvisna od vrednosti atributa *X1*. V podatkovni zbirki je 1000 primerov. Primeri so generirani po naslednjih pravilih:

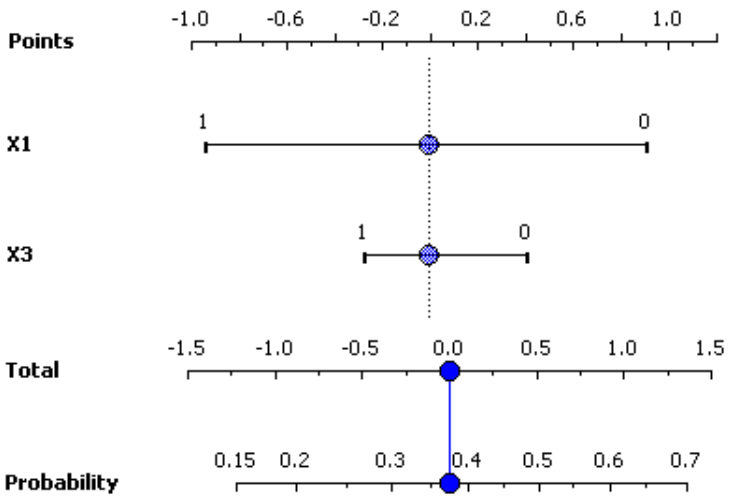
- V polovici primerov je X_1 enak 1
- Če je X_1 enak 1:
  - se razred ujema z vrednostjo X_1 v 50% primerov
  - X_2 se ujema z vrednostjo razreda v 80% primerov
  - X_3 se ujema z vrednostjo razreda v 80% primerov
- Če je X_1 enak 0:
  - se razred ujema z vrednostjo X_1 v 50% primerov
  - X_2 se ujema z vrednostjo razreda v 40% primerov
  - X_3 se ujema z vrednostjo razreda v 40% primerov

4.2.2 Eksperiment na umetni domeni UM1

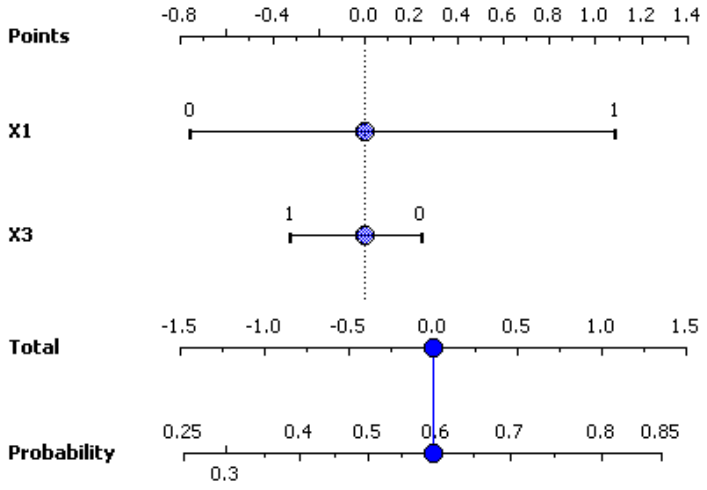


Slika 4.1: Primer nomograma naivnega Bayesovega klasifikatorja na umetnih domeni UM1

Poglejmo ali lahko te pogojne odvisnosti in neodvisnosti odkrijemo s pomočjo izbirnih nomogramov. Na sliki 4.1 je prikazan začetni nomogram naivnega Bayesovega klasifikatorja. Ko izberemo vrednost atributa *X2 = 1*, dobimo nomogram na sliki 4.2. Vrednosti atributa *X3* na nomogramski osi ostaneta enako oddaljeni in v enakem vrstnem redu. Če izberemo še drugo vrednost atributa *X2 = 0*, se zgodi podobno. Razdalja med vrednostima atributa *X3* se sicer nekoliko zmanjša, vendar ne bistveno. Nomogram je prikazan na sliki 4.3. Atributa *X2* in *X3* torej izgledata pogojno neodvisna. Povsem nekaj drugega se dogaja z atributom *X1*. Na nomogramu 4.2 se vrednosti močno oddaljita, v primerjavi z nomogramom 4.1. Ob izbrani vrednosti atributa *X2 = 0* pa celo zamenjata strani, kar je prikazano na sliki 4.3. Opazimo torej močno pogojno odvisnost med atributoma *X2* in *X1*. Če se osredotočimo na atribut *X3* in izbiramo vrednosti tega atributa, pridemo do podobnih ugotovitev. Ponovno opazimo pogojno neodvisnost med atributoma *X2* in *X1* ter pogojno odvisnost med atributoma *X3* in *X1*. Ugotovitve lahko preverimo še na atributu *X1*. Na sliki 4.4 je prikazan nomogram za izbiro *X1 = 0*. Na tem nomogramu sta nomogramskih osi obeh atributov, atributa *X2* in atributa *X3* bistveno ra-

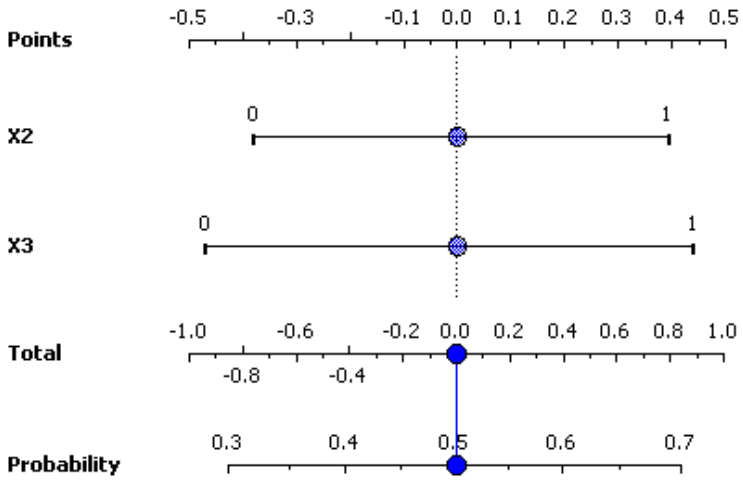


Slika 4.2: Primer nomograma naivnega Bayesovega klasifikatorja na umetnih domeni UM1 z izbrano vrednostjo atributa *X2 = 1*



Slika 4.3: Primer nomograma naivnega Bayesovega klasifikatorja na umetnih domeni UM1 z izbrano vrednostjo atributa *X2 = 0*

zlični od tistih na sliki 4.1. Z nadaljnjim izbiranjem vrednosti preostalih dveh atributov lahko le še potrdimo ugotovitev o neodvisnosti med tema atribu-



Slika 4.4: Primer nomograma naivnega Bayesovega klasifikatorja na umetnih domeni UM1 z izbrano vrednostjo atributa *X1 = 0*

toma. Z izbirnimi nomogramu smo torej na enostaven in hiter način odkrili pogojne (ne)odvisnosti med atributi.

4.3 Delovanje na domeni TITANIC

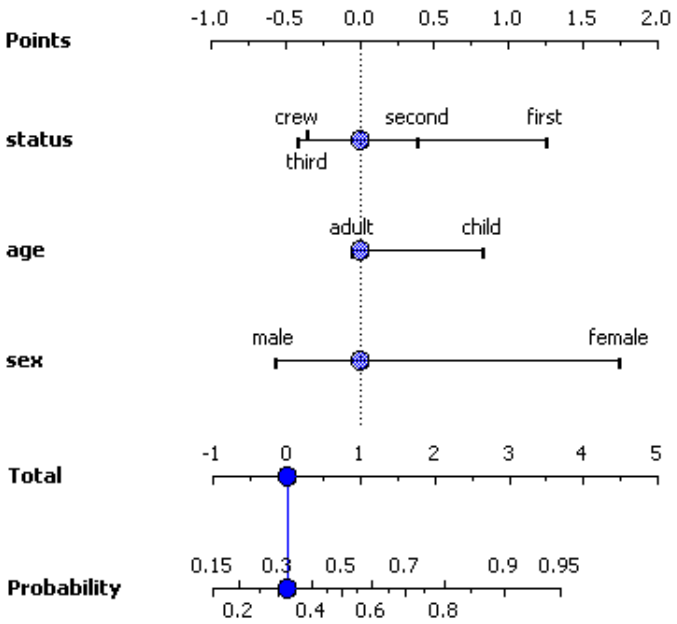
4.3.1 Opis domene TITANIC

V podatkih domene TITANIC je 2201 primer. Domena vsebuje tri diskretne atribute. Atribut *sex* z vrednostima *female* in *male*, atribut *age* z vrednostima *adult* in *child* ter atribut *status* z vrednostmi *first, second, third* in *crew*. Razred je diskretni atribut *survived* z vrednostima *yes* in *no*. Večinskemu razredu *no* pripada 67.7

4.3.2 Eksperiment na domeni TITANIC

Ta domena je vsekakor bolj kompleksna kot prejšnji dve.

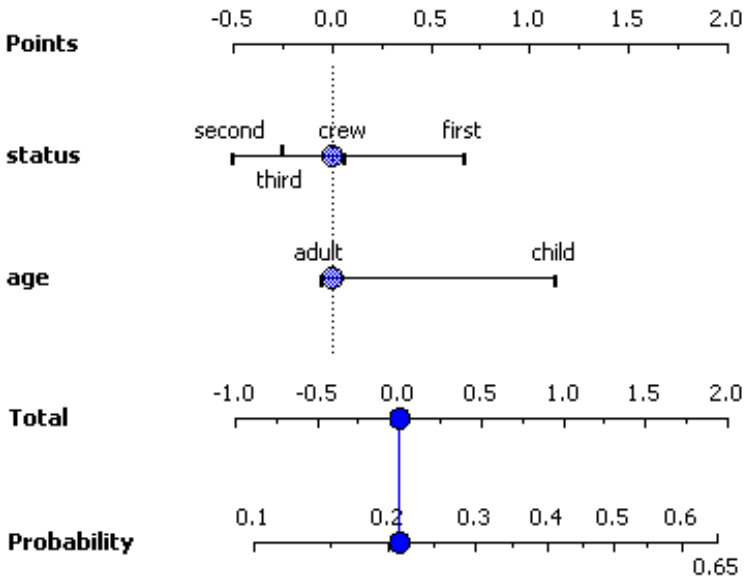
Na sliki 4.5 je nomogram naivnega Bayesovega klasifikatorja za podatke iz domene TITANIC s ciljnim razredom *survived = yes*. Kot lahko razberemo iz nomograma, imajo najvec moznosti za prezivetje otroci zenskega spola iz prvega razreda, najmanj pa odrasli moški iz posadke. Sedaj si podrobneje



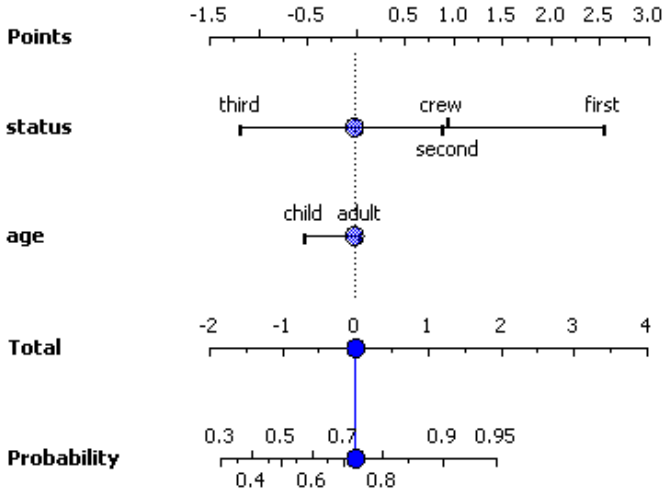
Slika 4.5: Nomogram naivnega Bayesovega klasifikatorja za podatke iz domene TITANIC

poglejmo moške. Izberemo vrednost atributa *sex* = *male*. Nov nomogram je prikazan na sliki 4.6.

Na tem nomogramu opazimo, da imajo pravzaprav moški iz drugega ali tretjega razreda manj možnosti za preživetje kot moški iz posadke. To se ne ujema s prejšnjimi ugotovitvami. To spremembo bi lahko povzročila visoka smrtnost med ženskami iz posadke. Na sliki 4.7 je nomogram za osebe ženskega spola. Zanimivo tudi tukaj se izkaže bivanje v tretjem razredu bolj smrtonoseno kot mesto med posadko. Ob podrobnejšem pregledu opazimo da sta drugi in tretji razred na nomogramih z ženskami in nomogramih z moškimi zamenjana. Torej je bilo med ženskami več smrtnih žrtev v tretjem razredu, med moškimi pa v drugem. Naivni Bayesov klasifikator računa le pogojne verjetnosti $P(r_k|v_i)$, torej verjetnost ciljnega razreda pri posameznih vrednostih atributa. V tem primeru se vpliv tretjega in drugega razreda, ki sta pri moških in ženskah obratna, izniči. Zato se za najbolj smrtonoseno izkaže mesto med posadko.
Tudi pri večvrednostnih atributih torej veljajo enake zakonitosti odkrivanja pogojnih odvisnosti med podatki.



Slika 4.6: Nomogram naivnega Bayesovega klasifikatorja za podatke iz domene TITANIC z izbrano vrednostjo atributa *sex = male*



Slika 4.7: Nomogram naivnega Bayesovega klasifikatorja za podatke titanic.tab z izbrano vrednostjo atributa *sex = female*

4.4 Delovanje na domeni RIBE

4.4.1 Opis domene RIBE

Domena RIBE je sestavljena iz treh diskretnih atributov in diskretnega razreda. Atribut *riba* lahko zavzame vrednosti *soška postrv, šarenka* ali *lipan* . Atribut *muha* lahko zavzame vrednosti *suha, ličinka* ali *potezanka*. Atribut *vreme* ima le dve možni vrednosti in sicer *oblačno* in *sončno*. Razred je atribut *prijem* z možnima vrednostima *da* in *ne*. Primeri opisujejo katero *ribo* je ribič lovil, katero vrsto *muhe* je uporabil in kakšno je bilo vreme *vreme* med ribolovom. Razred *prijem* nam pove, ali je bil pri ribolovu uspešen.

4.4.2 Eksperiment na domeni RIBE

V domeni RIBE je skritih veliko pogojnih odvisnosti. Lipan ima najraje sončno vreme in suho muho, soška postrv bo prijela le v oblačnem vremenu na potezanko, šarenka poje vse kar vidi, bolj aktivna pa je v oblačnih dneh. Vse to nam lahko povedo izkušeni ribiči. Ribič začetnik, ki še nima dovolj izkušenj, zna povedati le, da najpogosteje ujame šarenko, ribe najbolj prijemajo na ličinko in najraje v oblačnem vremenu. To je pravzaprav znanje, opisano na nomogramu s slike 3.1. Znanje izkušenega ribiča pa je znanje, ki ga pridobimo z uporabo izbirnih nomogramov. Del tega znanja je predstavljen na nomogramih s slik 3.2, 3.3 in 3.4.

Želimo si ujeti katerokoli ribo. Za nasvet vprašamo izkušenega ribiča, ki povzema znanje pridobljeno z izbirnimi nomogrami. Svetuje nam lov šarenke z ličinko v oblačnem vremenu. Enako nam svetuje tudi začetnik. Lovimo uspešno in obe napovedi se izkažeta za točni. Naslednji dan je sončen, a to nas ne odvrne od ribolova. Za nasvet spet povprašamo oba ribiča. Izkušen ribič svetuje lov lipana s suho muho, začetnik pa ponovno lov šarenke z ličinko. Še več, začetnik nam svetuje naj ne upoštevamo nasveta izkušenega ribiča, saj sicer ne bomo ujeli ribe. Poizkusimo oba načina in ujamemo lipana na suho muho. Napoved začetnika se izkaže za zgrešeno. Poglejmo še enkrat na sliko 3.1, kjer je nomogram, ki prikazuje njegovo znanje. Vrednosti atributov *riba=lipan* in *muha=suha* ne vplivata bistveno na verjetnost *prijema*. Vrednost atributa *vreme=sončno* pa zmanjša verjetnost *prijema*. To je razlog, da je začetnik napovedal neuspeh te strategije. Kombinacija vrednosti *riba=šarenka* in *muha=ličinka* pa dovolj pozitivno vplivata na verjetnost *prijema*, da je začetnik napovedal prijem kljub sončnemu vremenu. Težava je v tem, da začetnik našo namero, da bomo lovili v sončnem vremenu upošteva le

kot odločitev, ki nam zmanjša verjetnost prijema. Izkušen ribič premisli, kako se ribe v sončnem vremenu obnašajo in na kaj prijemajo. Izkušen ribič je torej v prednosti, saj se zna prilagajati situaciji. Prav tako so izbirni nomogrami v prednosti pred navadnimi, ker znajo upoštevajo odvisnosti med atributi. Ta prednost bi se lahko dobro izkazala pri klasifikaciji.

Poglavje 5

Klasifikacija s pomočjo izbirnih nomogramov

Idejo o izbirnih nomogramih bi lahko uporabili pri klasifikaciji. Podobno kot izkušnje pomagajo ribiču med ribolovom, lahko znanje pridobljeno s pomočjo izbirnih nomogramov pomaga pri klasifikaciji. Če izkušenemu ribiču povemo kdaj bomo lovili, katero vabo bomo uporabili in katero ribo bomo lovili, nam bo znal dokaj zanesljivo povedati ali bomo pri lovu uspešni. Medtem ko bo začetnik napovedoval precej nezanesljivo. Tako je zaradi pogojnih odvisnosti med tem kaj, na kaj in kdaj lovimo. Če teh odvisnosti ne bi bilo, bi oba napovedovala enako. Prav tako nam izbirni nomogrami ne bi razkrili nič novega, če bi bili vsi atributi pogojno neodvisni. Kako pa naj si z idejo o izbirnih nomogramih pomagamo pri klasifikaciji? Tako kot izkušen ribič, moramo upoštevati vrednosti atributov primera, ki bi ga radi klasificirali. V skrajnem primeru bi lahko klasificirali tako, da bi v izbirnih nomogramih izbrali prav vse vrednosti atributa, ki ustrezajo vrednostim v primeru. Tako bi dobili podprostor primerov, ki so identični našemu primeru po vseh atributih. Na podlagi teh primerov bi nato napovedali, kateremu razredu najverjetneje pripada naš primer.

Tak pristop pa prinese neželene posledice. Namesto izboljšanja klasifikacijske točnosti se ta pogosto zmanjša. Pogosto se tudi zgodi, da je v učni množici primerov, ki se s testnim primerom ujemajo v vrednosti vseh atributov, zelo malo ali jih celo ni. V tem primeru je napoved podana na podlagi zelo malo primerov in posledično nezanesljiva. To težavo lahko omilimo s širjenjem okolice.

5.1 Širjenje okolice

Če moramo klasificirati primer, ki je v učni množici redko zastopan, imamo težavo. Na podlagi le nekaj primerov težko sklepamo kateremu razredu pripada primer. Zato je smiselno pri napovedovanju upoštevati še nekaj najbolj podobnih primerov, torej razširiti okolico primera in na podlagi te okolice napovedati razred. Zaplete pa se pri podobnosti. Kaj je podobnost v diskretnem prostoru? Kaj je razdalja med primeri? Lahko bi uporabili Hammingovo razdaljo, ki meri v vrednosti koliko atributov se primera razlikujeta. Vendar ali je res vseeno po katerem atributu se razlikujeta? Vrnimo se k našemu izkušenemu ribiču. Recimo, da ga povprašamo ali lahko ujamemo soško postrv na suho muho v oblačnem vremenu on pa je to poizkusil le enkrat in to neuspešno. Ali naj nam na podlagi enkratne izkušnje odvrne, da je ne bomo ujeli? Tega seveda ne stori. Vendar na podlagi česa naj nam odgovori? Tudi šarenka spada v družino postrvi, morda lahko pri svoji odločitvi upošteva izkušnje s šarenko in suho muho v oblačnem vremenu. Ličinka je po velikosti podobna suhi muhi, morda lahko upošteva izkušnje pri lovu soške postrvi z ličinko v oblačnem vremenu. Lahko pa bi nam svetoval na podlagi lova soške postrvi na suho muho v sončnem vremenu. Ribič mora pravzaprav ugotoviti, kaj bolj vpliva na uspešnost lova, vrsta ribe, ki jo lovimo, tip muhe ki jo uporabljamo ali morda vreme. Po Hammingovi razdalji so vsi omenjeni primeri enako oddaljeni od obravnavanega primera, dvomim pa, da bi se ribič strinjal, da je vseeno kako se odloči. Če se odloči, da je najmanj pomembno vreme, bo njegov odgovor *ne*, saj soška postrv ne prijema na suho muho v sončnem vremenu. Če se odloči, da je najmanj pomembna vaba, bo njegov odgovor *ne*, saj soška postrv ne prijema na ličinko v oblačnem vremenu. Če pa se odloči, da je najmanj pomembna vrsta ribe, ki jo lovimo, bo odgovor *da*, saj šarenka rada zagrize v suho muho ko so na nebu oblaki. Našemu ribiču bi bilo najbolj všeč, če bi se lahko vzdržal odgovora, morda se tako kdaj počutijo tudi klasifikatorji?

5.1.1 Implementacija širjenja okolice

Poleg pomembnosti tega, po katerem atributu se primera razlikujeta, prihaja do razlik tudi med samimi vrednostmi atributa. Naš ribič je na podlagi izkušenj sklepal, da je soški postrvi bolj podobna šarenka kot pa lipan. Obe sta namreč postrvi. V tem elementu je ribič korak pred nami, saj mi načeloma ne vemo, kateri vrednosti posameznega diskretnega atributa sta si bolj podobni. Seveda tudi ni nujno, da je primer, ki se od obravnavanega primera razlikuje v vrednosti enega atributa, bolj podoben le-temu kot nek drug primer, ki ima različni

vrednosti dveh atributov. V zveznih domenah si je to lažje predstavljati. Tam sta si primera, ki se rahlo razlikujeta v vrednosti dveh atributov po evklidski razdalji bližje kot primera, ki se bistveno razlikujeta v vrednosti le enega.

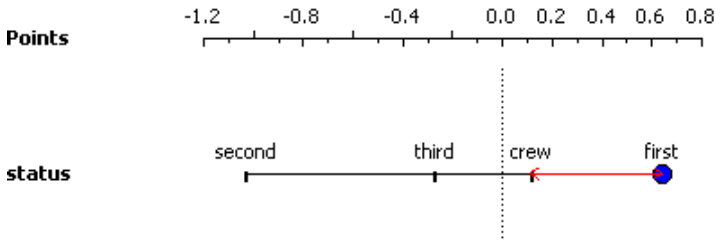
Primer na domeni TITANIC

Primer širjenja okolice bomo prikazali na domeni TITANIC, saj je ta domena dovolj kompleksna za nazoren prikaz delovanja. Težave z razdaljami v diskretnih domenah smo se lotili nomogrami Naivnega Bayesovega klasifikatorja. Razdaljo med primeri lahko namreč določimo na nomogramskih oseh nomograma. To naredimo tako, da v izbirnem Nomogramu izberemo vse vrednosti atributov, v katerih se primera ujemata, ostale pa pustimo nedoločene. Na nomogramskih oseh nato razberemo razdalje med vrednostima atributa, po katerem se primera razlikujeta. Končna razdalja je vsota teh razdalj. Nomogrami naivnega Bayesovega klasifikatorja namreč prikazujejo, kako posamezna vrednost atributa vpliva na verjetnost ciljnega razreda. Če se primera ujemata v vseh razen enem atributu in sta vrednosti atributa obeh primerov zelo blizu na nomogramski osi, potem imata ta dva primera zelo podoben vpliv na ciljni razred. Vpliv na ciljni razred pa je to kar nas zanima.

Vzemimo primer z naslednjimi vrednostmi atributov:

$$[status = first, sex = male, age = adult]$$

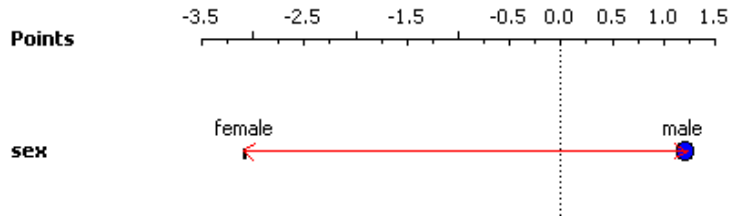
Na sliki 5.1 je prikazan nomogram z izbranima vrednostima *age=adult* in *sex=male*. Najbližja vrednost na sliki 5.1 je *status=crew* z razdaljo 0,52.



Slika 5.1: Razdalja med primeri z različnimi vrednostmi atributa *status*

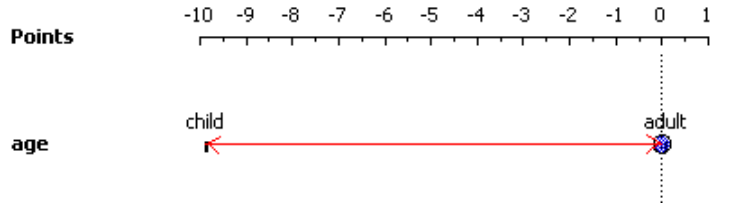
Na sliki 5.2 je prikazan nomogram z izbranima vrednostima atributov *status=first* in *age=adult*. Najbližja vrednost na sliki 5.2 je *sex=female* z razdaljo 4,29.

Na sliki 5.3 je prikazan nomogram z izbranima vrednostima atributov *sex=male*



Slika 5.2: Razdalja med primeri z različnimi vrednostmi atributa *sex*

in *status=first*. Najbližja vrednost na sliki 5.3 je *sex=female* z razdaljo 9,93.

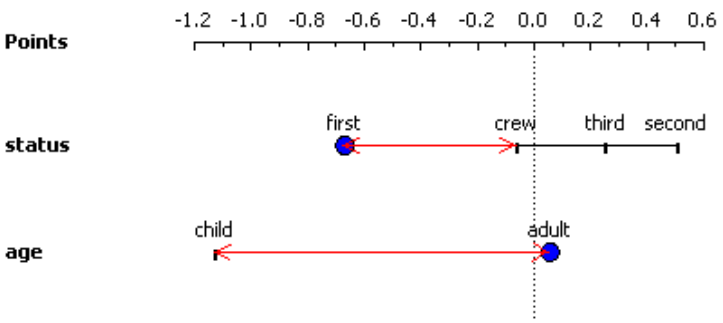


Slika 5.3: Razdalja med primeri z različnimi vrednostmi atributa *age*

Z izbirnimi nomogrami lahko tako določimo, kateri primeri imajo najbolj podoben vpliv na končni razred. Za najbolj podobne našemu primeru se izkažejo primeri, ki se od našega razlikujejo le po atributu *status*. Natančneje tisti primeri, ki imajo vrednost atributa *status* enako *crew*, v vrednostih ostalih atributov pa se ujemajo z našim primerom. Okolico torej razširimo tako da vanjo dodamo te primere.

Sedaj moramo paziti pri nadaljnjem širjenju. Upoštevati moramo, da smo že dodali primere z vrednostjo atributa *status* enako *crew*. Med kandidate za širitev poleg tistih, ki se od našega primera razlikujejo po enem atributu, dodamo še primere, ki imajo vrednost atributa *status* enako *crew* in se od našega primera razlikujejo še po vrednosti nekega drugega atributa. Razdalja se izračuna kot vsota razdalj na nomogramskih oseh med vrednostmi atributov našega primera in vrednostmi kandidata za širitev. Nomogram v takem primeru je prikazan na sliki 5.4. S takim postopkom omogočimo, da se okolica prej razširi s primeri, ki se od obravnavanega primera razlikujejo v vrednosti večih atributov, če so te primeri bolj podobni obravnavanemu primeru kot ostali primeri.

Skupna razdalja na nomogramu 5.4 je 1,79. To nam pove, da so primeri z vrednostmi atributov *status=crew, age=child, sex=male* bližje našemu primeru kot primeri, ki se od našega razlikujejo le po vrednosti enega atributa. Tako



Slika 5.4: Razdalja med primeri z vrednostjo različnimi vrednostmi atributa *age* in *status*

postopamo in dodajamo primere v okolico, dokler ne dosežemo zadostnega števila primerov. Na tako dobljeni okolici lahko nato izvajamo poljubne operacije.

5.2 Implementacija klasifikatorja

Zgoraj opisan način širjenja okolice privede do ideje o uporabnosti postopka pri klasifikaciji. Če naj bi bilo zgoraj opisano širjenje okolice boljše od širjenja s pomočjo Hammingove razdalje, bi ga lahko uporabili pri klasifikaciji podobni algoritmu kNN. Ta temelji ravno na lokalni klasifikaciji. Bližnje primere uteži glede na njihovo oddaljenost in nato uteženo upošteva vrednosti njihovih razredov pri klasifikaciji. V diskretnih domenah to počne s Hammingovo razdaljo. V tem razdelku bomo predstavili klasifikator, ki za iskanje okolice uporablja postopek opisan v prejšnjem razdelku.

5.2.1 Opis algoritma

Vhod: primer p, št. primerov v okolici $numEx$
Izhod: razred primera p Za vsako vrednost razreda naredimo sledeče:

- določimo obravnavano vrednost razreda za ciljno
- razširimo okolico primera p do $numEx$ primerov
- utežimo primere v okolici
- uporabimo uteženo vsoto za določitev verjetnosti ciljnega razreda

Nato kot razred vrnemo tisto vrednost razreda, ki ima največjo verjetnost. Uteževal sem glede na razdalje, izračunane enako kot v postopku širjenja okolice. Uteževal sem po formuli

$$e^{-t^2/s^2} \tag{5.1}$$

kjer je t razdalja od začetenga primera, vrednost s je določena tako, da ima najbolj oddaljen primer utež 0,001.

5.3 Rezultati

Algoritem smo implementirali v okolju Orange. Za testiranje klasifikatorja smo uporabljali 10-kratno prečno preverjanje. Primerjali smo ga z Bayesovim klasifikatorjem in klasifikatorjem kNN. Pri algoritmu kNN in algoritmu, ki ga obravnavamo v tem diplomskem delu je bilo uporabljeno enako število sosedov. S tem smo omogočili primerjanje kakovosti okolic posameznega algoritma. V tabeli 5.1 so podani rezultati testiranja klasifikatorjev.

	naš klasifikator	**naivni Bayes**	**kNN**
titanic	76.69%	77.83%	78.87%
zoo	87.12%	96.18%	91.18%
breast-cancer	63.33%	72.32%	74.83%
UM1-50	63.3%	70.00%	73.33%
UM1-1000	68.8%	64.6%	69%

Tabela 5.1: Rezultati prečnega preverjanja klasifikatorjev

Klasifikatorje smo testirali na domenah TITANIC, BREAST-CANCER, ZOO, umetni domeni UM1 s 50 primeri ter umetni domeni UM1 s 1000 primeri. Klasifikator se v praksi ni izkazal. Boljši je le od naivnega Bayesa na umetni domeni UM1 s 1000 primeri. Želeli bi si, da je boljši od metode kNN, saj bi to nakazovalo na kvalitetnejše iskanje okolice. Metodi kNN se približa le na umetni domeni UM1 s 1000 primeri.

Po natančni analizi algoritma in analizi okolic, ki jih le-ta zgradi in uporablja pri klasifikaciji, smo odkrili vzrok slabe klasifikacije. Težava je v širjenju okolice. V praksi pride do težav, ko je prostor redek. Torej ko je atributov veliko, primerov pa relativno malo, glede na število možnih kombinacij vrednosti atributov. V domeni Titanic imamo 2201 primer in le 16 možnih kombinacij vrednosti atributov (če ne upoštevamo razreda sicer jih je 32). Veliko primerov

se torej ponavlja. Domena UM1 ima tri binarne atribute in torej le $2^3 = 8$ možnih kombinacij atributov. Pri testiranju na 50 pimerih se primeri še ne ponavljajo tako pogosto, zato se naš klasifikator slabše izkaže. Ko imamo 1000 primerov pa se isti primeri pogosto ponavljajo. Klasifikator smo testirali tudi na domeni ZOO z dobrimi sto primeri in več tisoč možnimi kombinacijami vrednosti atributov. Do težav je prišlo, ker redkost prostora omeji zanesljivost izračunov pogojnih verjetnosti. Pogosto se namreč dogaja, da je le nekaj primerov ali celo ni primerov z določeno kombinacijo vrednosti atributov. Po zgoraj opisanem postopku moramo oceniti razdaljo med primeri na podlagi nomogramov naivnega Bayesovega klasifikatorja, kar je pri majhnem številu primerov zelo nezanesljivo. Okolica se v zelo redkih prostorih širi skorajda naključno. Omembe vreden je tudi podatek, da kNN z uporabo Hammingove razdalje ni imel posebnih težav na domenah kjer smo opazili težave našega klasifikatorja. Na vseh domenah se je izkazal bistveno bolje.

Poglavje 6

Zaključek

V tem diplomskem delu smo opisali izbirne nomograme, nov pristop k vizualizaciji naivnega Bayesovega klasifikatorja. Izkaže se, da izbirni nomogrami omogočajo natančnejšo analizo prostora in odkrivanje pogojnih odvisnosti med atributi. Od uporabnika ne zahtevajo poznavanja metod strojnega učenja, bistveno bolj pomembno je poznavanje domene. Zato so izbirni nomogrami primerni predvsem za eksperte, ki lahko z njimi analizirajo podatke in odkrijejo podprostore, kjer veljajo drugačne zakonitosti, kot tiste izpeljane iz osnovnega nomograma naivnega Bayesovega klasifikatorja. Nomogrami naivnega Bayesovega klasifikatorja so se v preteklosti že izkazali za razumljive nepoznavalcem področja in so pogosto uporabljan grafični pripomoček za razlago modelov strokovnjakom iz drugih področij. Izbirni nomogrami bi lahko, zaradi svoje interaktivnosti, pomenili še korak dlje v komunikaciji z eksperti. Razvili smo tudi metodo iskanja bližnje okolice primera, ki temelji na izbirnih nomogramih. Metodo iskanja okolice smo, žal neuspešno, uporabili za implementacijo klasifikatorja. Pri testiranju klasifikatorja smo odkrili nekatere pomanjkljivosti iskanja okolice. Težave so se pokazale predvsem v redkih prostorih, kjer se iskanje okolice izkaže za skorajda naključno in, kar je bistveno, precej slabše od iskanja okolice s Hammingovo razdaljo, ki jo uporablja kNN.

V nadaljnem delu bi se veljalo usmeriti v nadgradnjo vizualizacije s pomočjo izbirnih nomogramov. Za lažje primerjanje nomogramov pred in po izbiri vrednosti atributa, bi se lahko oba nomograma prikazovala hkrati, enega ob drugem. Smiselno bi bilo tudi nazorno označevati vrednosti atributov, katerih vplivi na razred se spreminjajo. Veliko prostora za nadaljne delo je tudi na področju širjenja okolice s predlagano metodo. Razdalja, ki bi nam v dikretnih domenah povedala kaj več kot le to po kolikih atributih se primera razlikujeta, bi bila uporabna na mnogih področjih strojnega učenja. Implemetacija klasi-

fikacije je bila neuspešna, vendar je še veliko prostora za izboljšave. Zanimivo bi bilo predvsem raziskovanje v smeri diskretnega odvajanja [11], s katerim bi lahko odkrivali domenske podprostore s podobnimi lastnostmi.

Slike

Tabele

Literatura

[1] J. Demšar, B. Zupan, "Orange: From Experimental Machine Learning to Interactive Data Mining," White Paper (www.ailab.si/orange), Faculty of Computer adn Information Science, University of Ljubljana, 2004

[2] N. Friedman, D. Greiger, M. Goldszmidt, "Bayesian network clasifiers," *Machine Learning*, št. 29, str. 131-163, 1997.

[3] A. Jakulin, I. Bratko, "Analyzing Attribute Dependencies," v zborniku *7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003)*, Cavtat, Hrvaška, sept. 2003

[4] A. Jakulin, I. Irish, *Bayesian Learning of Markov Network Structure*, Berlin: Springer, 2006, str. 198-209

[5] A. Jakulin, M. Možina, J.Demšar, I. Bratko, B. Zupan, "Nomograms for visualizing support vector machines," v zborniku *Proceedings of the eleventh ACM SIGKDD conference on Knowledge discovery in Data Mining*, New York, ZDA, 2005, str. 108-117.

[6] I. Kononenko, *Strojno učenje*, Ljubljana: Založba FE in FRI, 2005, pogl. 4.4

[7] I. Kononenko, *Strojno učenje*, Ljubljana: Založba FE in FRI, 2005, pogl. 10.1

[8] I. Kononenko, *Strojno učenje*, Ljubljana: Založba FE in FRI, 2005, pogl. 10.3

[9] J. Lubsen, J. Pool, E. van der Does, "A practical device for the application of a diagnostic or prognostic function," *Methods of Information in Medicine*, št. 17, str. 127-129, 1978.

[10] M. Možina, J. Demšar, M. Kattan, B. Zupan, "Nomograms for Visualization of Naive Bayesian Classifier," v zborniku *Knowledge Discovery in Databases: PKDD 2004*, Berlin, nov. 2004, str. 337-348.

[11] J. Žabkar, "Učenje kvalitativnih odvisnosti," doktorska disertacija, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana, Slovenija, 2010.