UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO
FAKULTETA ZA MATEMATIKO IN FIZIKO

Anže Starič

Pristopi strojnega učenja za tekmovanje UCSD Data Mining Contest

DIPLOMSKO DELO
NA INTERDISCIPLINARNEM UNIVERZITETNEM ŠTUDIJU

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2010

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

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/-a Anže Starič,

z vpisno številko 63060243,

sem avtor/-ica diplomskega dela z naslovom:

Pristopi strojnega učenja za tekmovanje UCSD Data Mining Contest

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom
  prof. dr. Blaža Zupana
- 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 7.9.2010 Podpis avtorja/-ice:

Zahvala

Zahvaljujem se mentorju prof. dr. Blažu Zupanu za predloge ter strokovno pomoč pri izdelavi diplomske naloge ter za vse znanje, ki se ga od njega prejel v času študija.

Prav tako se zahvaljujem sošolcema Gaji in Slavku, ki sta skozi vsa štiri leta skrbela za zapiske. Brez njiju bi bil študij težji.

Zahvaljujem se tudi podjetju Red Bull, ki mi je s svojimi izdelki omogočilo, da sem vse izpite opravil že v junijskem roku.

Iskrena hvala gre tudi mojim staršem in ženi Anji, saj me podpirajo in mi stojijo ob strani.

Kazalo

Povzetek

Sodelovanje na tekmovanjih iz strojnega učenja nas seznani z novimi problemskimi domenami in tipi problemov. Prisili nas, da raziščemo in spoznamo nove tehnike ter poiščemo inovativne pristope za reševanje problemov. Na tekmovanju UCSD Data Mining Contest smo se srečali s problemom razvrščanja uporabnikov glede na to, kdo bo najverjetneje postal kupec. V diplomskem delu je predstavljen razvoj tehnike, ki kar najbolje napove verjetnost, da bo določen uporabnik postal kupec. V sklopu razvoja smo preverili, kako se pri reševanju omenjenega problema obnesejo standardne metode strojnega učenja ter če je katera od značilk izrazito informativna. Naredili smo pregled razširitev naivnega Bayesovega klasifikatorja, ki izboljšujejo njegovo napovedno točnost ter najboljšo tudi implementirali z uporabo programskega paketa Orange. Preizkusili smo, kakšne napovedi daje naivni Bayesov klasifikator kot del ansambla klasifikatorjev ter kako na njegovo napovedno točnost vpliva diskretizacija podatkov. Rezultati so pokazali, da je določanje potencialnih kupcev za standardne tehnike strojnega učenja težak problem. Razširitve naivnega Bayesovega klasifikatorja so se odrezale bolje, najboljši rezultat je dosegel naivni Bayesov klasifikator v kombinaciji z ustrezno diskretizacijo podatkov. Dobljene ocene AUC so v primerjavi z ocenami AUC na drugih problemih, ki jih rešujemo s strojnim učenje, nizke, zato bi bilo potrebno pred uporabo razvite metode v praksi v podatkovno množico dodati še kakšno novo značilko.

Ključne besede:

strojno učenje
naivni Bayes
rangiranje
ansambli klasifikatorjev

Abstract

With participation in machine learning competitions we get acquainted with new problem domains and new types of problems. We are forced to look for and try out new techniques and search for innovative problem solving approaches. In UCSD Data Mining Contest, our task was to rank the ordering consumer pool according to who is most likely to become a customer of the retailer. In the following dissertation we have developed a technique for predicting the probability of a consumer becoming a customer of the retailer. Standard machine learning algorithms were evaluated and attribute analysis has been performed on the train dataset. In order to improve the score of standard algorithms review of methods that augment Naive Bayes for ranking has also been carried out and the most promising one has been implemented by using the Orange framework. We have also assessed the impact of data discretization on the Naive Bayes and evaluated ensemble techniques that combine the Naive Bayes Classifiers. Results show that ranking of potential customers is indeed a hard task for standard machine learning algorithms. Augmented Naive Bayes performed slightly better in terms of AUC, but the best results were produced using a combination of data discretization and standard Naive Bayes Classifier. AUC scores achieved were relatively low compared to scores achieved on other machine learning problems. This suggests that more attributes should be introduced into dataset before using this method in production environment.

Key words:

machine learning
Naive Bayes
ranking
ensemble techniques

Poglavje 1

Uvod

UCSD Data Mining Contest je tekmovanje, ki ga vsako leto organizira Univerza v San Diegu. Tema tekmovanja, katerega smo se udeležili, je določanje potencialnih novih kupcev na populaciji uporabnikov.

1.1 Odkrivanje bodočih kupcev

Spletne aplikacije že od svojih začetkov zbirajo podatke o uporabnikih. Z brezplačno registracijo uporabnik ponudniku zagotovi osnovne podatke, kot so ime, priimek, naslov in datum rojstva. Med uporabnikovim obiskovanjem strani se podatke dopolni z dodatnimi informacijami, kot so npr. povprečno trajanje obiska, kategorije zanimanja in podobne lastnosti, ki jih lahko ugotovimo z analizo uporabnikovih obiskov.

S tehnikami strojnega učenja lahko pomagamo ponudnikom iz že zbranih podatkov o uporabnikih določiti, kateri od uporabnikov bodo izdelek tudi kupili oziroma postali plačljivi uporabniki storitve. Ponudnik ima namreč neposredno korist le od takih uporabnikov. Pridobljeno znanje lahko ponudnik izkoristi pri izvajanju ciljnega oglaševanja, določanju prioritete prošenj uporabnika ter ostalih postopkih, pri katerih lahko več pozornosti nameni potencialnim kupcem in jih s tem prepriča v nakup.

1.2 Rangiranje

Vrsto strojnega učenja, pri kateri želimo vsakemu kupcu določiti verjetnost pripadnosti razredu, imenujemo rangiranje (angl. ranking). Od klasifikacije se

razlikuje v tem, da je poleg razreda, ki ga klasifikator priredi primeru, pomembno tudi, kako dobro oceno dobi posamezen primer v primerjavi z ostalimi.

V okviru diplomskega dela smo na praktičnem problemu preizkusili v času študija spoznane metode strojnega učenja. Raziskali smo, kako izboljšati naivni Bayesov klasifikator za reševanje problemov rangiranja ter najboljšo od tehnik implementirali z uporabo programskega paketa Orange [9]. Razvili smo tudi svojo napovedno tehniko in jo primerjali z obstoječimi.

Spoznali smo okolje xgrid, njegove prednosti in omejitve pri izvajanju testov na gruči računalnikov. Vsa koda, ki je bila razvita v okviru diplomske naloge, omogoča izvajanje na gruči, s čimer močno skrajšamo čas, potreben za izvajanje testov.

1.3 Struktura diplomskega dela

Poleg uvoda in zaključka diplomsko delo sestavljajo še tri osrednja poglavja. V drugem poglavju so predstavljene uporabljene tehnike za določanje tipa, diskretizacijo in ocenjevanje kvalitete značilk. Tretje poglavje opisuje uporabljene napovedne tehnike ter oceno AUC, s katero smo vrednotili kvaliteto napovedi. V četrtem poglavju je predstavljena podatkovna množica in opisan postopek razvoja ter testiranja napovednih tehnik. Podane so tudi ocene tehnik pridobljene z uporabo prečnega preverjanja na učni množici in uspešnost tehnik na testni množici. V zadnjem poglavju sledi komentar uporabljenih razširitev naivnega Bayesovega klasifikatorja ter predlogi za nadaljnje delo.

Poglavje 2

Analiza in izbor značilk

Objavljene podatkovne množice niso vsebovale podatkov o tipu in pomenu značilk, zato smo morali te podatke pridobiti sami. Pri tem smo uporabili kombinacijo računskih tehnik in ročnega pregledovanja.

2.1 Določanje vrste značilk

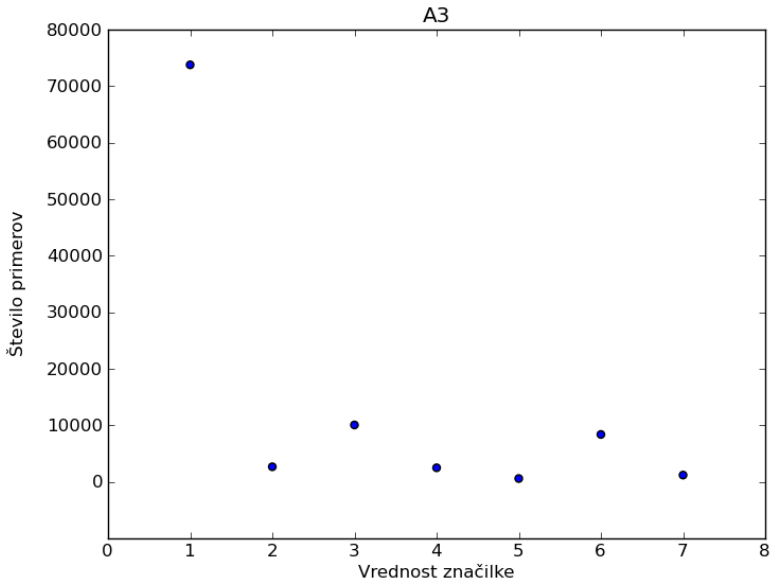
Pri atributnem učenju srečamo dve vrsti značilk, diskretne in zvezne. Vsaka vrsta služi predstavitvi različnega tipa podatkov, zato ju napovedne tehnike obravnavajo različno.

Diskretne značilke se uporabljajo za predstavitev podatkov, ki imajo končni nabor vrednosti. Vrednosti značilke se med seboj izključujejo, zato jih napovedne tehnike običajno obravnavajo tako, da ločeno obravnavajo vse možne vrednosti. Med vrednostmi običajno relacija urejenosti ne obstaja, zato je potrebno za tehnike, ki znajo delati le z zveznimi značilkami, diskretne značilke binarizirati.

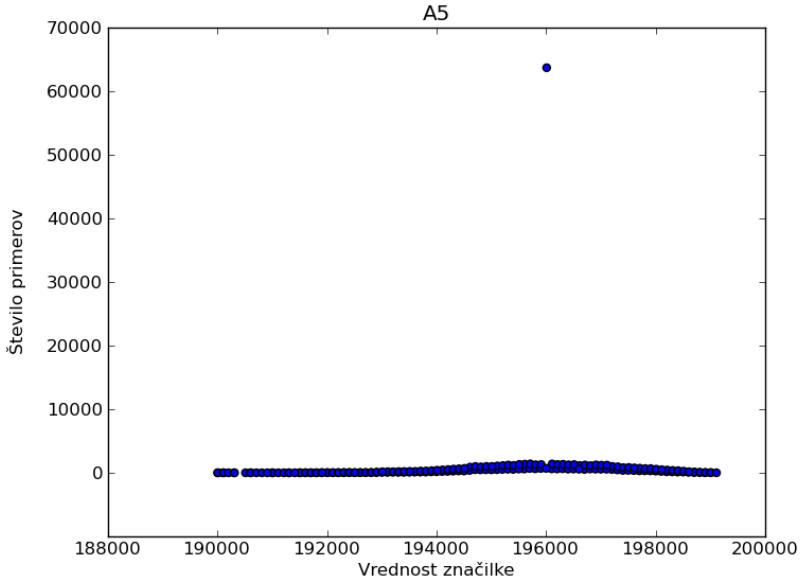
Binarizacija je postopek, pri katerem iz diskretne značilke D z vrednostmi $v_1, v_2, v_3, ..., v_k$ zgradimo k zveznih značilk $Z_1, Z_2, ..., Z_k$ z zalogo vrednosti $[0, 1]$. Značilka Z_i ima vrednost 1, če je vrednost značilke D enaka v_i, sicer ima vrednost 0.

Zvezne značilke se uporabljajo za predstavitev podatkov, za katere obstaja relacija urejenosti. Njihova zaloga vrednosti je običajno množica realnih števil. Napovedne tehnike, ki uporabljajo zvezne značilke, jih obravnavajo kot števila, bodisi kot vrednosti komponent vektorja ali parametre funkcij. Za tehnike, ki ne znajo obravnati zveznih značilk, je potrebno podatke diskretizirati.

V tekmovalnih podatkovnih množicah so bila kot zaloga vrednosti obeh vrst značilk uporabljena realna števila. Vrste značilk smo zato ločevali glede na razporeditev njihovih vrednosti. Za vsako značilko smo izrisali porazdelitev



Slika 2.1: Primer distribucije vrednosti diskretne značilke



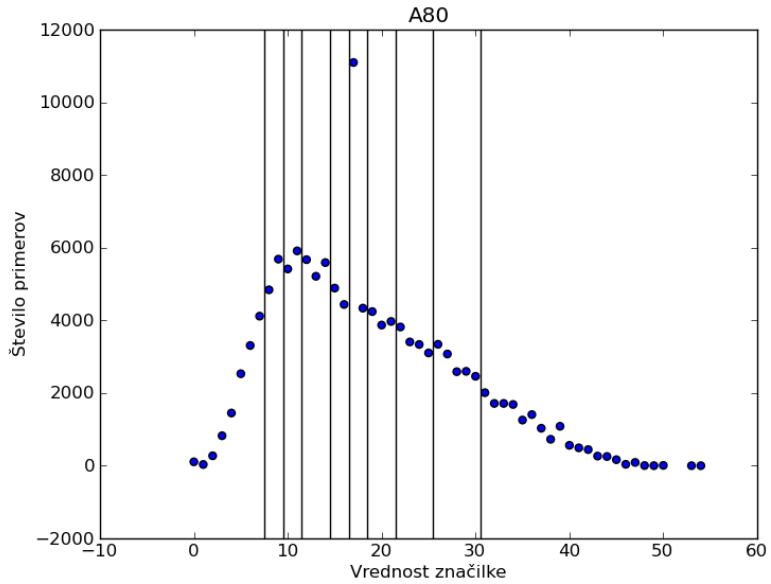
Slika 2.2: Primer distribucije vrednosti zvezne značilke

njenih vrednosti ter iz nje določili tip spremenljivke. Za diskretne smo označili značilke, ki so imele majhno število različnih vrednosti, bile pa so enakomerno razporejene po intervalu $[1, k]$, kjer je k število različnih vrednosti.

Na sliki 2.1 je primer porazdelitve vrednosti za diskretno značilko. Iz slike je razvidno, da ima značilka A3 sedem različnih vrednosti, ki pripadajo številom od ena do sedem. Slika 2.2 prikazuje porazdelitev vrednosti zvezne značilke. Opazimo lahko, da ima značilka A5 večje število različnih vrednosti, vrednosti pa se ne začnejo z ena.

2.2 Diskretizacija

Diskretizacija je postopek, pri katerem interval zaloge vrednosti zvezne značilke razbijemo na več podintervalov, ki predstavljajo diskretne vrednosti. Potrebujemo jo, ker nekatere tehnike strojnega učenja ne znajo obravnavati zveznih značilk.



Slika 2.3: Primer diskretizacije zvezne značilke

Raziskave [4] kažejo, da lahko s kombinacijo naivnega Bayesovega klasifikatorja in dobre diskretizacije pozitivno vplivamo na klasifikacijsko točnost. Raziskava pokaže, da so najprimernejše metode lena diskretizacija, proporcionalna delitev na k intervalov ter ne-disjunktna diskretizacija.

2.2.1 Diskretizacija na intervale z enakim številom primerov (EFD)

Diskretizacija na intervale z enakim številom primerov (angl. Equal Frequency Discretization) je osnovna diskretizacija, pri kateri iz željenega števila intervalov l izračunamo širino posameznega intervala s z uporabo enačbe:

$$s = \lfloor \frac{n}{s} \rfloor \tag{2.1}$$

Nato s prehodom zaloge vrednosti zvezne značilke določimo podintervale. Naslednji interval začnemo, ko prejšnji vsebuje več kot željeno število primerov. Iz slike 2.3 je razvidno, da imajo na ta način dobljeni intervali različne širine, vendar vsebujejo približno enako število primerov.

Dobra lastnost te vrste diskretizacije je, da so vsi intervali dobro zastopani. Slabost je, da dve vrednosti, ki sta si bili pri zvezni značilki blizu, obravnavamo drugače, če padeta v različna intervala. Vrednosti, ki sta na različnih robovih podintervala pa obravnavamo enako, čeprav sta si bili v zvezni značilki bolj oddaljeni.

2.2.2 Sorazmerna delitev na k intervalov (PKID)

PKID [5] (angl. Proportional k-Interval Discretization) se obnaša enako kot diskretizacija na intervale z enakim številom primerov, le da tu število intervalov ni znano vnaprej. Za izračun števila primerov v posameznem intervalu s in števila intervalov l uporabimo naslednji enačbi:

$$s \times l = n \tag{2.2}$$

$$s = l \tag{2.3}$$

Število primerov v posameznem intervalu je torej $s = \sqrt{n}$, kjer je n število vseh učnih primerov. Uporaba PKID naj bi izrazito zmanjšala napako pri klasifikaciji z uporabo naivnega Bayesovega klasifikatorja.

2.2.3 Lena diskretizacija

Pri leni diskretizaciji (angl. Lazy Discretization) z dejansko diskretizacijo počakamo do trenutka, ko uvrščamo nov primer. Takrat za vsako značilko,

za katero je potrebna diskretizacija, določimo podinterval, tako da se vrednost značilke nahaja na sredini intervala, interval pa vsebuje toliko primerov, kot pri EFD s parametrom $t = 10$

Lena diskretizacija naj bi izboljšala napovedno točnost naivnega Bayesovega klasifikatorja, vendar zaradi velike porabe pomnilnika (celotno učno množico si je potrebno zapomniti do trenutka klasifikacije) in računske zahtevnosti (za uvrščanje vsakega primera moramo diskretizirati učno množico in oceniti pogojne verjetnosti) ni primerna za uporabo na velikih podatkovnih množicah.

2.2.4 Ne-disjunktna diskretizacija (NDD)

NDD [6] (angl. Non-Disjoint Discretization) temelji na ideji, da je zamenjava vrednosti z intervalom, na katerem se nahaja, boljša, če se vrednost nahaja na sredini, kot če se nahaja robu intervala. Diskretizacija razdeli zalogo vrednosti na atomarne intervale, nato pa nad njimi zgradi podintervale, tako da posamezen podinterval vsebuje tri sosednje atomarne intervale. Vsaka vrednost se tako nahaja v treh različnih intervalih, kar upoštevamo pri ocenjevanju pogojnih verjetnosti.

Naivni Bayesov klasifikator, implementiran v paketu Orange, za oceno pogojnih verjetnosti za zvezne značilke uporablja tehniko, ki je podobna NDD. Za vsako vrednost značilke izračuna pogojno verjetnost pripadnosti razredu, nato pa z uporabo metode LOESS dobljene pogojne verjetnosti zgladi. S tem doseže enak učinek, kot če bi pri NDD za širino atomarnega intervala izbrali en primer, širina podintervala pa je enaka širini okna, ki ga pri metodi LOESS uporabljamo za glajenje.

2.3 Ocenjevanje značilk

Izmed vseh značilk smo določili bolj pomembne z uporabo dveh metod za oceno. Pri izboru, katere vrednosti ocen označujejo dobre značilke, smo uporabili permutacijski test.

2.3.1 Informacijski prispevek

Informacijski prispevek (angl. Information Gain) je mera, ki za ocenjevanje koristnosti značilk uporablja entropijo. Definiran je kot prispevana informacija značilke za določitev vrednosti razreda.

Pri računanju informacijskega prispevka značilke mera predpostavi medsebojno neodvisnost značilk, zato ne zmore prepoznati značilk, ki so pomembne

le v kombinaciji, ne pa kot posamezne. Mera je definirana le za diskretne atribute, zato smo zanjo podatke diskretizirali.

2.3.2 ReliefF

ReliefF [2] je izboljšana različica mere RELIEF. RELIEF je nekratkovidna mera, kar pomeni, da značilke ocenjuje v odvisnosti od ostalih značilk. To omogoča, da kot pomembne prepozna tudi značilke, ki so močno odvisne med seboj.

RELIEF za vsak primer v množici poišče najbližji primer z istim razredom in najbližji primer z različnim razredom. Če je vrednost značilke različna za primera z različnim razredom in enaka za primera z istim razredom, se ocena pomembnosti značilke poveča, v nasprotnem primeru pa zmanjša.

ReliefF je različica osnovne mere, ki namesto enega najbližjega primera vsake vrste uporabi k primerov. To jo naredi bolj odporno na šumne podatke. Pri izvajanju smo uporabili $k = 5$, namesto vseh pa je algoritem naključno izbral 100 primerov ter glede na njih ocenil značilke.

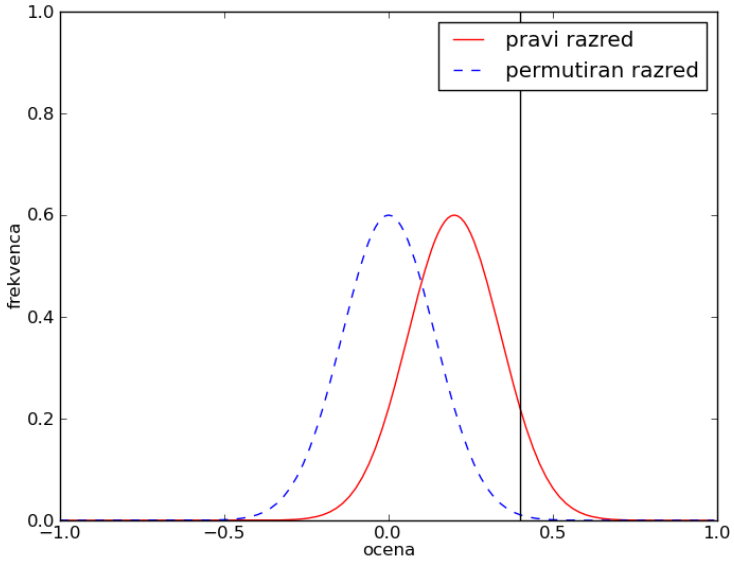
2.3.3 Permutacijski test

Pri ocenjevanju pomembnosti značilk mera za ocenjevanje značilkam dodeli številsko oceno njihove pomembnosti. Ocene težko interpretiramo, z njimi lahko primerjamo pomembnost dveh značilk, ali razvrstimo značilke od bolj do manj pomembnih. Ne moremo pa samo iz vrednosti ocen določiti, kje je meja med pomembnimi in nepomembnimi značilkami. Pri določanju meje si pomagamo s permutacijskim testom.

Permutacijski test je postopek, pri katerem primerjamo izračunane ocene pomembnosti značilk s porazdelitvijo ocen na naključnih podatkih. Za izračun te porazdelitve na učnih primerih permutiramo vrednosti razredne spremenljivke in na spremenjeni množici izračunamo ocene pomembnosti značilk. S permutacijo razredne spremenljivke smo odstranili odvisnost vrednosti značilk od vrednosti razreda, zato pričakujemo, da nobena značilka ne bo označena kot pomembna.

Rezultat permutacijskega testa sta porazdelitvi ocen mere za napoved kvalitete atributov na pravih in naključnih podatkih. Primer takih porazdelitev je prikazan na sliki 2.4. Z rdečo je narisana porazdelitev vrednosti ocene na pravih podatkih, z modro pa porazdelitev na naključnih podatkih.

Črna navpična črta označuje mejo kvalitetnih značilk. Postavljena je tako, da je levo od nje 99 % vrednosti ocen značilk na naključnih podatkih. Značilke,



Slika 2.4: Primer permutacijskega testa

ki so na pravih podatkih dosegle ocene, ki so višje od tega praga, jih najverjetneje niso dosegle zaradi naključja, zato jih označimo kot pomembne.

Poglavje 3

Učne metode

3.1 Naivni Bayesov klasifikator

Naivni Bayesov klasifikator napoveduje verjetnost pripadnosti posameznemu razredu z uporabo pogojne odvisnosti od vrednosti značilk. Pri tem predpostavlja medsebojno pogojno neodvisnost značilk glede na razred.

Naj r predstavlja iskani razred, x pa vektor vrednosti značilk $(x_1, x_2, ..., x_n)$.

Verjetnost pripadnosti iskanemu razredu lahko z uporabo Bayesovega pravila zapišemo kot

$$P(r|x) = P(r)\frac{P(x|r)}{P(x)}. \tag{3.1}$$

Ob predpostavki, da so vrednosti značilk glede na razred neodvisne

$$P(x|r) = \prod_{i=1}^{n} P(x_i|r) \tag{3.2}$$

dobimo

$$P(r|x) = \frac{P(r)}{P(x)} \prod_{i=1}^{n} P(x_i|r). \tag{3.3}$$

S ponovno uporabo Bayesovega pravila

$$P(x_i|r) = P(x_i)\frac{P(r|x_i)}{P(r)} \tag{3.4}$$

dobimo

$$P(r|x) = P(r)\frac{\prod_{i=1}^{n} P(x_i)}{P(x)} \prod_{i=1}^{n} \frac{P(r|x_i)}{P(r)}. \tag{3.5}$$

Faktor

$$\frac{\prod_{i=1}^{n} P(x_i)}{P(x)} \tag{3.6}$$

je neodvisen od razreda, zato ga lahko izpustimo. Tako dobimo končno formulo

$$P(r|x) = P(r) \prod_{i=1}^{n} \frac{P(r|x_i)}{P(r)}. \tag{3.7}$$

Pri učenju klasifikatorja je potrebno iz učne množice aproksimirati verjetnosti na desni strani enačbe. Te so apriorna verjetnost razreda in pogojne verjetnosti razreda glede na vrednosti posameznih značilk. Za uvrščanje novega primera s pomočjo enačbe 3.7 izračunamo verjetnost pripadnosti posameznemu razredu pri podanih vrednostih značilk. Klasifikator vrne razred, katerega verjetnost je največja.

3.2 Delno naivne tehnike

Ker je predpostavka neodvisnosti značilk na realnih podatkih pogosto kršena, verjetnosti, ki jih napove naivni Bayesov klasifikator, ne ustrezajo dejanskim. Za uporabo na problemih rangiranja so zato bolj primerne tehnike, ki vsaj delno upoštevajo tudi odvisnosti med značilkami [7].

3.2.1 NBTree

NBTree [1] je tehnika, ki združuje naivni Bayesov klasifikator in odločitvena drevesa. Z uporabo drevesne strukture razdeli učne primere na več podmnožic ter na vsaki podmnožici zgradi naivni Bayesov klasifikator. V notranjih vozliščih se primeri delijo na enak način kot pri odločitvenih drevesih, vsakemu vozlišču pripada ena značilka, primer glede na njeno vrednost dodelimo ustreznemu nasledniku. Vsak list vsebuje naivni Bayesov klasifikator, ki uvršča primere. Primeri so v liste razporejeni glede na vrednosti značilk, ki se nahajajo v vozliščih na poti od korena do posameznega lista. Iz tega sledi, da imajo te značilke v vseh primerih v listu enako vrednost, zato jih pri gradnji naivnega Bayesovega klasifikatorja ne upoštevamo.

Pomanjkljivost tehnike NBTree je v tem, da poiskuša zgraditi eno drevo, ki bi bilo primerno za uvrščanje vseh primerov. Vendar je zelo verjetno, da tako drevo ne bo dosegalo dobrih rezultatov v listih, v katerih je zastopanost primerov iz učne množice majhna. Naivni Bayesov klasifikator, ki bo zgrajen v takem listu, bo zgrajen na premajhnem številu primerov, zato bodo njegove napovedi slabe.

3.2.2 Leno učenje bayesovskih pravil

Leno učenje bayesovskih pravil [8] (angl. Lazy Learning of Bayesian Rules) poiskuša pomanjkljivost tehnike NBTree odpraviti z lenim pristopom. Pri tej tehniki gradnjo drevesa odložimo do trenutka, ko moramo uvrstiti nov primer. Za vsak primer, ki ga uvrščamo, zgradimo le tisto vejo, na kateri se vrednosti značilk v vozliščih ujemajo z vrednostmi značilk primera. Nato na učnih primerih v listu naučimo naivni Bayesov klasifikator ter z njim uvrstimo testni primer.

Pri lenem učenju bayesovskih pravil ima list, v katerega spada nov primer, vedno dovolj primerov, saj za to poskrbimo ob gradnji drevesa. Napovedi naivnih Bayesovih klasifikatorjev v listih so zato boljše, kar pomeni, da so tudi napovedne točnosti celotnega modela boljše. Pomanjkljivost, ki močno omeji uporabnost tehnike, je časovna zahtevnost uvrščanja novih primerov. Za vsak primer je potrebno zgraditi novo drevo, zato je tehnika primerna le za probleme, kjer je potrebno uvrstiti malo testnih primerov.

3.2.3 AODE

Tehnika AODE [3] (angl. Aggregating One-Dependence Estimators) se omeji na obravnavo paroma odvisnih značilk.
Iz definicije pogojne verjetnosti sledi

$$P(r|x) = \frac{P(r \wedge x)}{P(x)} \tag{3.8}$$

$$\propto P(r \wedge x). \tag{3.9}$$

Iz pravila produkta sledi, da za vsako vrednost značilke x_i velja

$$P(R = r \wedge x) = P(r \wedge x_i) P(x|r, x_i). \tag{3.10}$$

Ker velja za vse vrednosti, velja tudi za povprečje, torej

$$P(r \wedge x) = \frac{\sum_{i=1}^{n} P(r \wedge x_i) P(x|r, x_i)}{n}. \tag{3.11}$$

Pri predpostavki, da so vrednosti značilke razen x_i med seboj neodvisne, lahko zapišemo kot

$$P(r \wedge x) = \frac{\sum_{i=1}^{n} P(r \wedge x_i) \prod_{j=1}^{n} P(x_j|r \wedge x_i)}{n}. \tag{3.12}$$

Imenovalec enačbe 3.12 je neodvisen od razreda, zato ga lahko izpustimo. Če želimo namesto verjetnosti produkta dobiti pogojno verjetnost, števec še normiramo po vseh razredih in dobimo končno obliko enačbe:

$$P(r|x) = \frac{\sum_{i=1}^{n} P(r \wedge x_i) \prod_{j=1}^{n} P(x_j|r \wedge x_i)}{\sum_{r' \in R} \sum_{i=1}^{n} P(r' \wedge x_i) \prod_{j=1}^{n} P(x_j|r' \wedge x_i)}. \tag{3.13}$$

Klasifikator lahko zgradimo na dva načina. Prvi način je, da iz učne množice aproksimiramo pogojne verjetnosti na desni strani enačbe 3.13 in jih shranimo v tabelo. Pri uvrščanju novega primera iz tabele preberemo ustrezne pogojne verjetnosti ter z uporabo enačbe 3.13 izračunamo pogojne verjetnosti pripadnosti posameznim razredom.

Pri drugem načinu za vsako značilko A_i izdelamo notranji napovedni model, ki upošteva, da so vrednosti značilk odvisne od razreda in značilke A_i. Primer takega modela je NBTree, pri katerem je drevo sestavljeno le iz korena, ki podatke deli glede na vrednost značilke A_i in listov. Tako dobimo n različnih modelov. Pri uvrščanju novega primera z vsakim od notranjih modelov izračunamo verjetnosti pripadnosti posameznim razredom ter vrnemo povprečje napovedi modelov.

Druga tehnika je še posebej primerna za vzporedno izvajanje, saj so gradnje napovednih modelov med seboj neodvisne. Tudi pri uvrščanju se lahko poslužimo vzporednosti, saj lahko notranji modeli računajo verjetnosti neodvisno.

3.3 Ansambli klasifikatorjev

V strojnem učenju poznamo tehnike, ki namesto enega samega klasifikatorja na učni množici izgradijo več modelov. Za uvrščanje testnega primera vsak model izdela svojo napoved, nato pa z glasovanjem med modeli izberemo razred, v katerega bomo primer uvrstili.

3.3.1 Bagging

Bagging (angl. Bootstrap Aggregating) je tehnika, pri kateri pred učenjem modela uporabimo metodo stremena (angl. Bootstrap). Metoda stremena je tehnika razmnoževanja učnih primerov, pri kateri iz učne množice z n primeri izberemo s ponavljanjem n primerov. V povprečju se v tako generirani množici pojavi 63,2 % primerov iz prvotne množice.

Z metodo stremena kreiramo t množic ter na vsaki naučimo napovedni model. Modeli se med seboj razlikujejo, saj so bili zgrajeni na različnih podatkovnih množicah. Pri uvrščanju novih primerov napovemo razred z vsakim od modelov, napovedi pa združimo z glasovanjem. Če nas zanima verjetnost pripadnosti razredu, namesto razreda, ki je prejel največ glasov vrnemo frekvenčno porazdelitev glasov med posamezne razrede.

3.3.2 Nevedni Bayesov klasifikator (NeB)

Izraz nevedni Bayes (angl. Ignorant Bayes) izhaja iz fraze blažena nevednost (angl. Ignorance is bliss). V določenih primerih, nam večje število uporabljenih značilk ne poveča napovedne točnosti napovednega modela. Pri naivnem Bayesovem klasifikatorju to drži, ko so vrednosti značilk med seboj odvisne.

Pri gradnji nevednega Bayesovega klasifikatorja zgradimo t podmnožic značilk, tako da v vsako podmnožico izberemo značilko z verjetnostjo p. Pričakovana velikost podmnožice je $\frac{1}{p}$. Na vsaki podmnožici nato naučimo naivni Bayesov klasifikator.

Nov primer uvrstimo tako, da zanj z vsakim modelom napovemo razred. Za združevanje napovedi imamo več možnosti. Lahko jih združimo z glasovanjem, verjetnost pripadnosti razredu je v tem primeru delež modelov, ki so glasovali v prid temu modelu. Drugi način združevanja je, da napovedane verjetnosti pripadnosti razredom povprečimo ter vrnemo razred, ki ima največjo verjetnost. Glede na način združevanja napovedi smo implementirali dva klasifikatorja. NeB-glasovanje združuje napovedi modelov z glasovanjem, NeB-povprečenje pa s povprečenjem verjetnosti.

Tudi če v učni množici obstajajo značilke, katerih vrednosti so med seboj odvisne glede na dani razred, lahko generirane podmnožice še vedno ustrezajo predpostavki o medsebojni neodvisnosti značilk. Neodvisnost velja, ko odvisne značilke niso izbrane v isto podmnožico. Na tako generiranih podmnožicah pričakujemo, da bodo napovedi naivnega Bayesovega klasifikatorja dobre.

3.4 Ocena AUC

Ocena AUC (angl. Area Under Curve) je ocena izpeljana iz krivulje ROC (angl. Receiver Operating Characteristic). Krivuljo ROC rišemo za binarni klasifikacijski problem. Vodoravna os predstavlja delež napačno uvrščenih negativnih primerov (angl. FP rate), navpična os pa delež pravilno uvrščenih pozitivnih primerov (angl. TP rate).

Za klasifikator, ki za vsak uvrščen primer vrne verjetnost pripadnosti pozitivnemu razredu, lahko za vsak prag med razredoma na grafu določimo točko. Krivulja ROC je konveksna ovojnica tako dobljenih točk. Ocena AUC je definirana kot ploščina pod ROC krivuljo. Vrednost ocene AUC predstavlja verjetnost, da bo klasifikator pravilno razločil med pozitivnim in negativnim primerom (pozitivnemu primeru bo določil večjo verjetnost).

Poglavje 4

Eksperimentalno ovrednotenje

4.1 Predstavitev in obdelava podatkov

Podatke so bili sestavljeni iz dveh podatkovnih množic. Učna množica je vsebovala 130475, testna pa 86691 primerov. Vsak učni primer je bil podan z vrednostmi 334 neoznačenih numeričnih značilk. Razredna spremenljivka je bila binarna, 90 % primerov je pripadalo negativnemu, 10 % primerov pa pozitivnemu razredu.

Za testiranje obstoječih in implementacijo novih napovednih tehnik smo uporabili programski paket Orange. Meritve smo izvajali na gruči računalnikov s pomočjo okolja xgrid, ki služi delitvi nalog med več računalniki.

Tipe značilk smo določili s postopkom, opisanim v razdelku 2.1. Za vse značilke smo izrisali porazdelitev njihovih vrednosti ter iz njih določili, za katero vrsto značilke gre. Tabela značilk in ugotovljenih tipov se nahaja v dodatku A.

Učno in testno množico smo pretvorili v Orange podatkovni zapis. Ob pretvorbi smo poleg celotne učne množice zgradili še manjši vzorec, v katerega smo vsak učni primer uvrstili z verjetnostjo 1 %. Vzorec smo uporabljali pri razvoju novih metod, saj je testiranje na celotni učni množici vzelo precej časa. Uporabili smo ga tudi za oceno, koliko časa se bo neka metoda izvajala na celotni učni množici.

4.2 Izbor metod za klasifikacijo

Na vzorcu podatkov smo izvedli 10-kratno prečno preverjanje klasifikatorjev, ki so vključeni v programski paket Orange (odločitveno drevo, naivni

Bayes, metoda podpornih vektorjev, linearna regresija). Glede na čas izvajanja prečnega preverjanja na vzorcu podatkov smo ocenili čas izvajanja na celotni učni množici ter nato preverjanje ponovili še na njej. Čas izvajanja 10-kratnega prečnega preverjanja na celotni učni množici in dobljene ocene AUC smo zbrani v tabeli 4.1.

Tabela 4.1: Rezultati prečnega preverjanja vgrajenih metod

Metoda	Čas izvajanja	AUC
odločitveno drevo	6 ur	0,5129
naivni bayes	5 min	0,5978
linearna regresija	6 min	0,5884
SVM	>24ur (izvajanje je bilo prekinjeno)	

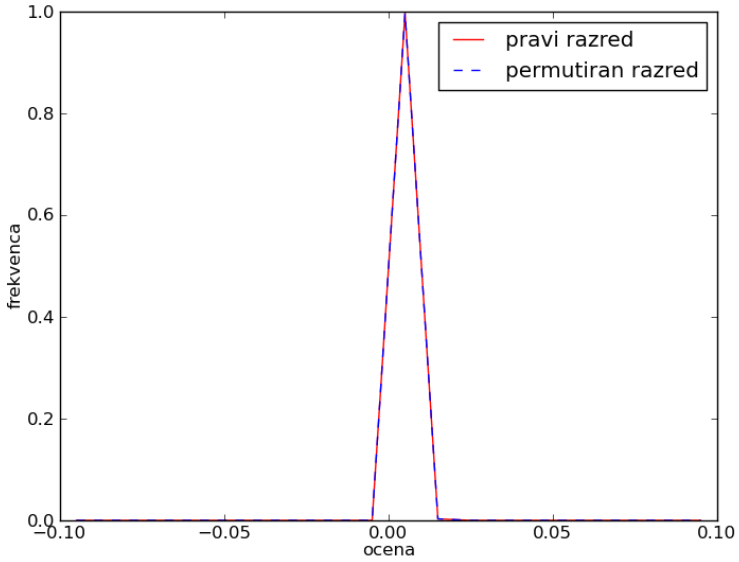
Zaradi počasnega izvajanja in slabega rezultata odločitvenih dreves in metode podpornih vektorjev nismo vključili v nadaljnje meritve. Linearna regresija je v paketu Orange implementirana z uporabo knjižnjice LibLinear, ki je hitra in robustna. Postopka gradnje modela zato nismo mogli spreminjati, zaradi česar nismo imeli veliko prostora za izboljševanje rezultata. V nadaljnjem testiranju smo se osredotočili na izboljšanje AUC ocene naivnega Bayesovega klasifikatorja.

4.3 Ocenjevanje značilk

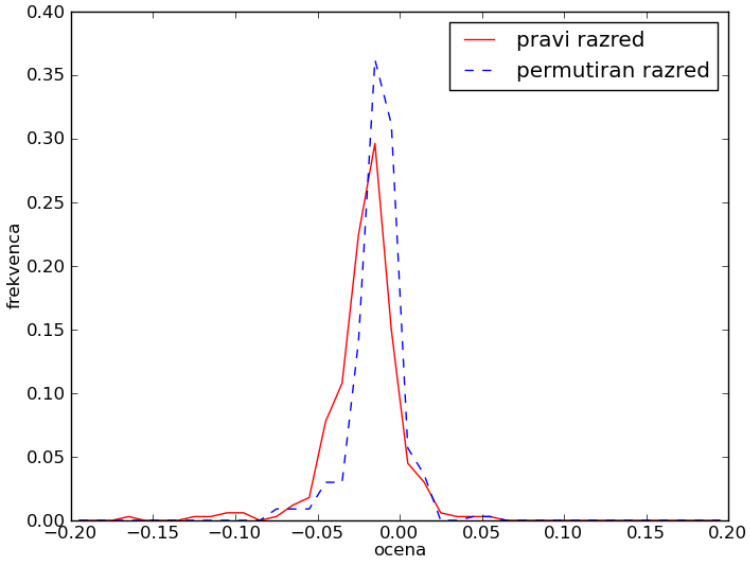
Izmerili smo informacijski prispevek značilk (2.3.1) in njihovo oceno z mero ReliefF (2.3.2). Sliki 4.1 in 4.2 prikazujeta rezultat permutacijskega testa za obe meri. Distribucija ocen informacijskega prispevka se popolnoma ujema z distribucijo na naključnih podatkih. Na obeh so namreč vse značilke bile ocenjene z 0,01 ali manj. Distribucija ocen mere ReliefF je glede na distribucijo podatkov s permutiranim razredom zamaknjena v levo, kar pomeni, da so bile značilke ocenjene kot manj pomembne. Ker z metodama izbora značilk nismo našli nobene izrazito dobre značilke smo z analizo posameznih značilk prenehali.

4.4 Ansambli klasifikatorjev

Oceno AUC smo poiskusili izboljšati z uporabo ansamblov klasifikatorjev. Uporabili smo metodo bagging na naivnem Bayesovem klasifikatorju ter metodo



Slika 4.1: Permutacijski test za mero Information Gain



Slika 4.2: Permutacijski test za mero ReliefF

nevedni Bayes (3.3.2), katero smo razvili sami. Uporabili smo obe različici metode.

Vrednost parametra t smo pri vseh metodah določili z 10-kratnim prečnim preverjanjem iz množice $t \in \{10, 15, 20, 25, 30\}$. Ocene AUC za različne vrednosti parametra t so prikazane v tabeli 4.2.

Tabela 4.2: Določanje parametra t

	AUC		
t	bagging	NevedniBayes-G	NevedniBayes-P
10	0,5985	0,5970	0,6051
15	0,5999	0,6006	0,6074
20	0,6007	0,6035	0,6061
25	**0,6013**	0,6030	**0,6097**
30	**0,6013**	**0,6041**	0,6074

4.5 AODE

Implementirali smo tehniko AODE, ki obravnava morebitne odvisnosti parov atributov. Iz ocene časa izvajanja na vzorcu učne množice smo ugotovili, da bo izvajanje na celotni učni množici vzelo preveč časa. Klasifikator smo zato razbili na več delnih klasifikatorjev, od katerih je vsak obravnaval odvisnost od ene značilke in dobljene delne klasifikatorje vzporedno naučili na gruči. Tudi uvrščanje testnih primerov smo pognali vzporedno, vsak delni klasifikator je izdelal svojo datoteko napovednih verjetnosti. Te smo v naslednjem koraku združili in izračunali verjetnosti pripadnosti pozitivnemu razredu.

4.6 Diskretizacija podatkov

Na podatkih smo uporabili različne tehnike diskretizacije. Na vsaki množici smo z uporabo 10-kratnega prečnega preverjanja testirali naivni Bayesov klasifikator. V tabeli 4.3 so zbrani rezultati preverjanja.

Iz rezultatov opazimo, da je diskretizacija z metodo PKID res boljša od diskretizacije EFD z vidika napovedne točnosti. Vendar pa pri oceni AUC doseže slabši rezultat. Vseeno je diskretizacija v obeh primerih izboljšala napovedno točnost in AUC glede na testiranje na originalni učni množici.

Tabela 4.3: Uspešnost naivnega Bayesa na diskretiziranih množicah

Vrsta diskretizacije	AUC	CA
Brez diskretizacije	0,5978	0,4795
EFD, k=10	0,6214	0,6515
PKID	0,6175	0,6608

4.7 Rezultati

Z uporabo opisanih tehnik smo naučili klasifikatorje še na celi učni množici. Z vsakim smo nato napovedali verjetnost pripadnosti pozitivnemu razredu za vsak primer iz testne množice in rezultate oddali na strežnik tekmovanja. Tabela 4.4 vsebuje rezultate klasifikatorjev na testni množici.

Tabela 4.4: AUC ocena algoritmov na testnih podatkih

Metoda	AUC
naivni Bayesov klasifikator	0,6149
AODE	0,6296
Bagging naivnih Bayesovih klasifikatorjev	0,6022
NeB-glasovanje	0,6234
NeB-povprečenje	0,6296
naivni Bayes na diskretiziranih podatkih (EFD, k=10)	**0,6372**

4.8 Diskusija

Obe implementirani metodi, ki poskušata upoštevati odvisnosti med atributi tako AODE kot nevedni Bayesov klasifikator, dosežeta boljšo oceno AUC od naivnega Bayesovega klasifikatorja. Nekoliko presenetljivo je dejstvo, da se na testni množici najbolje obnese naivni Bayesov klasifikator, naučen na diskretizirani učni množici. Razlog za to se morda skriva v tem, da so vrednosti zveznih značilk med seboj odvisne. Ko značilke diskretiziramo na 10 intervalov, dobimo nove vrednosti, ki med seboj niso tako močno odvisne, zato so napovedi naivnega Bayesovega klasifikatorja na njih boljše.

Poglavje 5

Zaključek

V diplomskem delu smo se ukvarjali z napovedovanjem verjetnosti, da primer pripada pozitivnemu razredu na dani podatkovni množici. V reševanju tega problema smo preizkusili obstoječe klasifikatorje, ki so vgrajeni v programski paket Orange, dodatno pa smo implementirali še dva klasifikatorja, ki izhajata iz naivnega Bayesovega klasifikatorja, a poiskušata upoštevati tudi pogojne odvisnosti med vrednostmi značilk. Ocenili smo tudi kvaliteto značilk in preizkusili nekaj različnih vrst diskretizacije.

Med izdelavo smo spoznali kar nekaj novih tehnik, ki razširjajo naivni Bayesov klasifikator. Naučili smo se rokovanja z večjimi podatkovnimi množicami ter dela z okoljem xgrid, s katerim smo lahko meritve izvajali vzporedno.

V diplomskem delu smo se osredotočili predvsem na tehnike, ki izhajajo iz naivnega Bayesovega klasifikatorja, oziroma so z njim povezane. V okviru nadaljnjega dela na tem področju bi lahko preverili, kako se pri reševanju problema napovedovanja potencialnih kupcev obnesejo metode, ki so sorodne linearni regresiji, le da namesto klasifikacijske točnosti optimizirajo oceno AUC.

Dodatek A

Tipi značilk

Ime značilke	Število vrednosti	Tip značilke	Ime značilke	Število vrednosti	Tip značilke
A1	3	diskretna	A26	17	diskretna
A2	42	zvezna	A27	72	zvezna
A3	8	diskretna	A28	5200	zvezna
A4	42	zvezna	A29	11	diskretna
A5	173	zvezna	A30	56	zvezna
A6	4	diskretna	A31	77	diskretna
A7	5	diskretna	A32	3131	zvezna
A8	10	diskretna	A33	31	diskretna
A9	10	diskretna	A34	11	diskretna
A10	4	diskretna	A35	27	diskretna
A11	197	zvezna	A36	29	zvezna
A12	9	diskretna	A37	30	zvezna
A13	22	diskretna	A38	62	zvezna
A14	134	zvezna	A39	63	zvezna
A15	3	diskretna	A40	449	zvezna
A16	197	zvezna	A41	432	zvezna
A17	402	zvezna	A42	452	zvezna
A18	334	zvezna	A43	380	zvezna
A19	381	zvezna	A44	433	zvezna
A20	24	diskretna	A45	21	diskretna
A21	1313	zvezna	A46	1144	zvezna
A22	35	zvezna	A47	940	zvezna
A23	1186	zvezna	A48	1148	zvezna
A24	42	zvezna	A49	1209	zvezna
A25	3360	zvezna	A50	1155	zvezna

Ime značilke	Število vrednosti	Tip značilke	Ime značilke	Število vrednosti	Tip značilke
A51	803	zvezna	A91	13	diskretna
A52	608	zvezna	A92	8	diskretna
A53	985	zvezna	A93	41	zvezna
A54	638	zvezna	A94	49	diskretna
A55	37	zvezna	A95	50	zvezna
A56	74	zvezna	A96	39	zvezna
A57	37	zvezna	A97	41	zvezna
A58	51	zvezna	A98	17	diskretna
A59	63	zvezna	A99	11	diskretna
A60	8044	zvezna	A100	21	diskretna
A61	9624	zvezna	A101	10	diskretna
A62	13124	zvezna	A102	70	zvezna
A63	88	zvezna	A103	62	zvezna
A64	100	zvezna	A104	91	zvezna
A65	66	zvezna	A105	54	zvezna
A66	71	zvezna	A106	67	zvezna
A67	27	diskretna	A107	27	diskretna
A68	100	zvezna	A108	100	zvezna
A69	92	zvezna	A109	34	diskretna
A70	72	zvezna	A110	47	zvezna
A71	100	zvezna	A111	94	zvezna
A72	56	diskretna	A112	9	diskretna
A73	37	diskretna	A113	39	zvezna
A74	78	diskretna	A114	8	diskretna
A75	86	zvezna	A115	31	diskretna
A76	42	zvezna	A116	59	zvezna
A77	38	zvezna	A117	35	zvezna
A78	99	zvezna	A118	46	zvezna
A79	23	diskretna	A119	100	zvezna
A80	53	zvezna	A120	66	zvezna
A81	65	zvezna	A121	98	zvezna
A82	57	zvezna	A122	58	zvezna
A83	51	zvezna	A123	46	zvezna
A84	44	zvezna	A124	37	zvezna
A85	63	zvezna	A125	53	zvezna
A86	69	zvezna	A126	35	zvezna
A87	41	zvezna	A127	34	zvezna
A88	34	diskretna	A128	53	zvezna
A89	30	diskretna	A129	36	zvezna
A90	69	zvezna	A130	37	zvezna

Ime značilke	Število vrednosti	Tip značilke	Ime značilke	Število vrednosti	Tip značilke
A131	23	diskretna	A171	49	zvezna
A132	57	zvezna	A172	62	zvezna
A133	31	diskretna	A173	54	zvezna
A134	21	diskretna	A174	46	zvezna
A135	40	zvezna	A175	37	zvezna
A136	35	zvezna	A176	75	zvezna
A137	24	zvezna	A177	95	zvezna
A138	20	zvezna	A178	94	zvezna
A139	98	zvezna	A179	90	zvezna
A140	78	zvezna	A180	85	zvezna
A141	41	zvezna	A181	46	zvezna
A142	9	diskretna	A182	94	zvezna
A143	57	zvezna	A183	26	zvezna
A144	76	zvezna	A184	20	zvezna
A145	30	zvezna	A185	59	zvezna
A146	22	zvezna	A186	40	zvezna
A147	68	zvezna	A187	41	zvezna
A148	41	zvezna	A188	45	zvezna
A149	41	zvezna	A189	47	zvezna
A150	49	zvezna	A190	34	zvezna
A151	53	zvezna	A191	74	zvezna
A152	36	zvezna	A192	91	zvezna
A153	64	zvezna	A193	35	zvezna
A154	57	zvezna	A194	92	zvezna
A155	43	zvezna	A195	63	zvezna
A156	55	zvezna	A196	84	zvezna
A157	21	zvezna	A197	87	zvezna
A158	80	zvezna	A198	98	zvezna
A159	36	zvezna	A199	60	zvezna
A160	83	zvezna	A200	100	zvezna
A161	13	diskretna	A201	84	zvezna
A162	95	zvezna	A202	58	zvezna
A163	24	diskretna	A203	99	zvezna
A164	77	zvezna	A204	74	zvezna
A165	14	diskretna	A205	100	zvezna
A166	86	zvezna	A206	100	zvezna
A167	18	zvezna	A207	100	zvezna
A168	22	zvezna	A208	94	zvezna
A169	16	zvezna	A209	93	zvezna
A170	100	zvezna	A210	100	zvezna

Ime značilke	Število vrednosti	Tip značilke	Ime značilke	Število vrednosti	Tip značilke
A211	100	zvezna	A251	19	diskretna
A212	85	zvezna	A252	22	diskretna
A213	46	zvezna	A253	39	zvezna
A214	64	zvezna	A254	48	zvezna
A215	21	zvezna	A255	39	zvezna
A216	85	zvezna	A256	55	zvezna
A217	83	zvezna	A257	45	zvezna
A218	54	zvezna	A258	28	zvezna
A219	23	zvezna	A259	41	zvezna
A220	56	zvezna	A260	53	zvezna
A221	70	zvezna	A261	47	zvezna
A222	95	zvezna	A262	65	zvezna
A223	10	zvezna	A263	48	zvezna
A224	94	zvezna	A264	77	zvezna
A225	41	zvezna	A265	54	zvezna
A226	46	zvezna	A266	54	zvezna
A227	86	zvezna	A267	42	zvezna
A228	95	zvezna	A268	58	zvezna
A229	84	zvezna	A269	100	zvezna
A230	112	zvezna	A270	100	zvezna
A231	89	zvezna	A271	100	zvezna
A232	76	zvezna	A272	100	zvezna
A233	83	zvezna	A273	90	zvezna
A234	91	zvezna	A274	53	zvezna
A235	86	zvezna	A275	100	zvezna
A236	72	zvezna	A276	65	zvezna
A237	88	zvezna	A277	97	zvezna
A238	62	zvezna	A278	96	zvezna
A239	100	zvezna	A279	22	zvezna
A240	97	zvezna	A280	100	zvezna
A241	100	zvezna	A281	100	zvezna
A242	91	zvezna	A282	100	zvezna
A243	9	diskretna	A283	40	zvezna
A244	39	zvezna	A284	75	zvezna
A245	84	zvezna	A285	42	zvezna
A246	52	zvezna	A286	61	zvezna
A247	43	zvezna	A287	20	zvezna
A248	37	zvezna	A288	45	zvezna
A249	18	diskretna	A289	100	zvezna
A250	33	zvezna	A290	60	zvezna

Ime značilke	Število vrednosti	Tip značilke
A291	100	zvezna
A292	29	zvezna
A293	90	zvezna
A294	96	zvezna
A295	100	zvezna
A296	100	zvezna
A297	100	zvezna
A298	32	zvezna
A299	92	zvezna
A300	48	zvezna
A301	35	zvezna
A302	57	zvezna
A303	65	zvezna
A304	67	zvezna
A305	38	zvezna
A306	100	zvezna
A307	100	zvezna
A308	97	zvezna
A309	26	zvezna
A310	88	zvezna
A311	46	zvezna
A312	91	zvezna
A313	11	zvezna
A314	39	zvezna
A315	7	diskretna
A316	21	zvezna
A317	75	zvezna
A318	31	zvezna
A319	15	zvezna
A320	88	zvezna
A321	94	zvezna
A322	72	zvezna
A323	95	zvezna
A324	100	zvezna
A325	84	zvezna
A326	77	zvezna
A327	100	zvezna
A328	37	zvezna
A329	31	zvezna
A330	33	zvezna
A331	90	zvezna
A332	75	zvezna
A333	785	zvezna
A334	653	zvezna
Razred	2	diskretna

Literatura

[1] R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," v zborniku *Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, USA, 1996.

[2] M. Robnik-Šikonja, I. Kononenko "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, št. 53, zv. 1, str. 23-69, 2003.

[3] G. I. Webb, J. R. Boughton, Z. Wang "Not so naive bayes: Aggregating one-dependence estimators," *Machine Learning*, št. 51, zv. 1, str. 5-24, 2002.

[4] Y. Yang, G. I. Webb, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers," v zborniku *Pacific Rim Knowledge Acquisition Workshop*, Tokyo, Japan, 2002, str. 159-173.

[5] Y. Yang, G. I. Webb, "Proportional k-interval discretization for naive-Bayes classifiers," v zborniku *Twelfth European Conference on Machine Learning*, 2001, str. 564-575.

[6] Y. Yang, G. I. Webb, "Non-disjoint discretization for Naive-Bayes classifiers," v zborniku *Nineteenth International Conference on Machine Learning*, 2002, str. 666-673.

[7] H. Zhang, L. Jiang, J. Su, "Augmenting Naive Bayes for Ranking," v zborniku 22^{nd} *International Conference on Machine Learning*, Bonn, Germany, 2005, str. 1027.

[8] Z. Zheng, G. I. Webb, "Lazy Learning of Bayesian Rules," *Machine Learning*, št. 41, zv. 1, str. 53-84, 2000.

[9] (2010) Programski paket Orange. Dostopno na: http://www.ailab.si/orange