UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matej Pičulin

DISKRETIZACIJA ŠTEVILSKIH ATRIBUTOV Z RAZVRŠČANJEM

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: prof. dr. Marko Robnik-Šikonja

Ljubljana, 2011

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: 01719/2010
Datum: 15.12.2010

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

Kandidat: **MATEJ PIČULIN**

Naslov: **DISKRETIZACIJA ŠTEVILSKIH ATRIBUTOV Z RAZVRŠČANJEM**
CLUSTERING-BASED DISCRETIZATION OF NUMERIC ATTRIBUTES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Nekateri algoritmi strojnega učenja in podatkovnega rudarjenja potrebujejo za svoje delovanje diskretizirane atribute. Diskretizacija jeza trenutno uveljavljene pristope težavna v problemih z močno odvisnimi atributi. Preizkusite idejo diskretizacije, kjer najprej izvedemo postopek razvrščanja, ki nam lahko predlaga naravne meje med skupinami podobnih primerov, potem pa te delitve izkoristimo za diskretizacijo posameznih atributov. Pri tem je potrebno odgovoriti na številna odprta vprašanja: kakšen algoritem razvrščanja uporabiti, koliko skupin naj vrne, kje postaviti meje med skupinami, koliko mej naj ima posamezen atribut in kako te meje določiti izmed predlaganih kandidatov. Odgovore na ta vprašanja poiščite z vizualizacijo in evalvacijo na umetnih in realnih domenah.

Mentor:
prof. dr. Marko Robnik Šikonja

Dekan:
prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

I Z J A V A O A V T O R S T V U

diplomskega dela

Spodaj podpisani Matej Pičulin,

z vpisno številko 63030182,

sem avtor diplomskega dela z naslovom:

Diskretizacija številskih atributov z razvrščanjem

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom

  prof. dr. Marka Robnik-Šikonje

- 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 15.3.2011 Podpis avtorja: Pičulin Matej

Zahvala

Zahvaljujem se mentorju, prof. dr. Marku Robnik-Šikonji, za pomoč in usmeritve pri izdelavi diplomskega dela.

Zahvaljujem se tudi svoji družini, ki mi je omogočila študij, ter vsem ostalim, ki so me podpirali v času študija.

Kazalo

Seznam uporabljenih kratic in simbolov

AUC ang. Area Under the ROC Curve - ploščina pod krivuljo ROC
CA ang. Classification Accuracy - klasificijska točnost
CD ang. Critical Diffrence - kritična razlika
CV ang. Cross Validation - prečno preverjanje
k-NN ang. K-Nearest Neighbors - k-najbližjih sosedov
MDL ang. Minimum Description Length - princip najkrajšega opisa
RF ang. Random Forest - naključni gozdovi
ROC ang. Reciever Operating Characteristic - ROC krivulja
SD ang. Standard Deviation - standardni odklon
UCI ang. University of California, Irvine

Povzetek

V nalogi predlagamo novo metodo za diskretizacijo, ki skuša uporabiti metode razvrščanja za določitev mej. Uporabili smo dve dobro znani metodi razvrščanja in sicer razvrščanje k-means ter hierarhično razvrščanje.

Diskretizacija je dobro znan in težak problem v strojnem učenju in podatkovnem rudarjenju, posebno pri atributih, ki so med seboj močno odvisni. Ker veliko obstoječih metod ne upošteva odvisnosti med atributi, smo poizkušali razviti algoritem, ki bo poskušal odvisnosti implicitno zaznati s pomočjo razvrščanja.

Na začetku predstavimo nekaj obstoječih metod diskretizacije in klasifikatorje, ki so uporabljeni v nalogi. Predstavimo idejo diskretizacije z razvrščanjem, pri čemer skušamo odgovoriti na osnovna vprašanja: katero vrsto razvrščanja uporabiti, koliko skupin je potrebnih, kako skupine glasujejo za meje ter kako iz kandidatov za meje izločiti končne meje. Metode smo preverili s testi na umetnih domenah z močnimi odvisnostmi in na realnih domenah. Uporabimo nekaj različic diskretizacije z razvrščanjem. Pokažemo, da se s to metodo da rešiti nekatere primere z močnimi odvisnostmi. Na koncu predstavimo še nekaj možnih izboljšav in idej za nadaljnje delo.

Ključne besede:

diskretizacija, razvrščanje, stojno učenje, številski atributi

Abstract

We propose a new method for discretization, which uses clustering to determine candidate boundaries. We use two well-known clustering methods: k-means clustering and hierarchical clustering.

Discretization is well-know and difficult problem in machine learning and data mining, especially for strongly dependent attributes. Most existing methods do not take dependencies into account, therefore we develop an algorithm, which will finds dependencies implicitly with the help of clustering.

First we present some known discretization methods and classification algorithms, which we use in the presentation. We present the idea of clustering-based discretization and try to answer the following questions: which clustering method to use, how many clusters do we need, how do clusters vote for boundaries and how to choose final boundaries from candidates. We extensively test the approach on artificial domains with strong dependencies and on real domains. We test several variations of cluster-based discretization and show the methods can solve some cases with strongly dependent attributes. Finally, we suggest possible improvements and extensions of the work.

Keywords:

discretization, clustering, machine learning, numeric attributes

Poglavje 1

1 Uvod

Veliko algoritmov strojnega učenja in podatkovnega rudarjenja za svoje delovanje potrebuje diskretizirane atribute. Veliko podatkov, ki jih dobimo z merjenem ima številsko (numerično oz. zvezno) obliko kot so na primer višina, dolžina, vsebnost holesterola v krvi, temperatura itd. Naloga diskretizacije je, da iz številskega atributa naredi diskretnega pri tem pa meje postavi tako, da je izgubljene čim manj informacije. Na primer, višino razdelimo na diskretne vrednosti: visok, srednji in nizek. To je v splošnem težaven problem, še posebno, če imamo več atributov, ki so med seboj močno odvisni.

Predstavimo novo metodo za diskretizacijo, s katero poizkušamo poiskati odvisnosti med številskimi atributi. Meje za diskretizacijo poizkušamo dobiti s pomočjo razvrščanja (ang. clustering oz. slovensko rojenje), ki ima to lastnost, da združuje primere, ki so si med seboj podobni.

V 2. poglavju predstavimo uporabljene klasifikatorje in mere ocenjevanja klasifikatorjev. Obstoječe diskretizacije uporabljene v tem delu so opisane v 3. poglavju. V 4. poglavju opišemo idejo diskretizacije z razvrščanjem skupaj z možnimi različicami in uporabljenimi rešitvami. V 5. poglavju preverimo, kako se diskretizacija z razvrščanjem izkaže na umetnih in realnih domenah. V zaključku podamo glavne ugotovitve ter predlagamo razširitve in ideje za nadaljnje delo.

Poglavje 2

2 Klasifikatorji in vrednotenje modelov

Za primerjavo algoritmov smo uporabili dve meri za ocenjevanje učenja in sicer klasificijsko točnost in AUC ter naslednje štiri klasifikatorje:

- naivni Bayes,
- odločitveno drevo,
- naključni gozdovi,
- k-najbližjih sosedov.

Metode so povzete po [3], kjer si lahko ogledamo tudi podrobnejši opis.

2.1 Uporabljeni klasifikatorji

2.1.1 Naivni Bayesov klasifikator

Naivni Bayesov klasifikator (ang. naive Bayes classifier) je eden od najbolj uporabljenih klasifikatorjev in je izpeljan iz Bayesovega pravila:

$$P(r_k \mid V) = P(r_k)\frac{P(V \mid r_k)}{P(V)}$$

Kjer je $P(r_k)$, $k = 1...n_0$ apriorna verjetnost razredov, $P(V)$ apriorna verjetnost primera z atributnim opisom V in $P(V|r_k)$ pogojna verjetnost primera z atributnim opisom v pri danem razredu r_k.

Naivni Bayes predpostavlja pogojno neodvisnost vrednosti različnih atributov pri danem razredu, zaradi česar je tudi dobil ime naivni. Kljub svoji naivnosti se v praksi pokaže kot dober

klasifikator tudi kadar neodvisnost med atributi ne drži popolnoma. Veliko slabše deluje pri močnih odvisnostih atributov.

Končna formula za uporabo Bayesovega klasifikatorja dobljena iz Bayesovega pravila ob upoštevanju neodvisnosti atributov je:

$$P(r_k \mid V) = P(r_k)\prod_{i=1}^{a}\frac{P(r_k \mid v_i)}{P(r_k)},$$

kjer je $P(r_k|v_i)$, $k = 1...n_0$ pogojna verjetnosti razreda pri dani vrednosti v_i atributa A_i, $i = 1...a$.

2.1.2 Odločitveno drevo

Odločitvena drevesa so pogosto uporabljena metoda za klasifikacijo v strojnem učenju. Metoda zgradi drevo, v katerem predstavljajo notranja vozlišča atribute, veje drevesa ustrezajo podmnožicam vrednosti atributov in listi razredom. Dobra lastnost odločitvenih dreves je, da so klasifikatorji razumljivi, ker se da drevesa vizualizirati in iz njih razbrati zakonitosti, ki se pojavljajo v modelu.

Osnovni algoritem za učenje odločitvenih dreves je naslednji:

Če je izpolnjen ustavitveni pogoj,
    potem postavi list, ki vključuje vse učne primere;
    sicer
        izberi “najboljši” atribut A_i;
        označi naslednike z vrednostmi atributa A_i;
        za vsako vrednost V_j atributa A_i ponovi:
        rekurzivno zgradi poddrevo z ustrezno podmnožico učnih primerov;

Ustavitveni pogoji so lahko naslednji:

- dovolj čista učna množica,
- premalo učnih primerov za zanesljivo nadaljevanje gradnje drevesa,
- zmanjkalo je dobrih atributov.

V testih v 4. poglavju je za izbiro najboljšega atributa uporabljen algoritem MDL, druge pogoste možnosti so še informacijski prispevek, razmerje informacijskega prispevka, Gini-indeks in ReliefF. V našem primeru je uporabljena binarizacija, kar pomeni, da se notranje vozlišče vedno razdeli na dva dela. Za ustavitveni pogoj je izbrana 100% čistost ali vsaj 2 elementa v listu drevesa.

V nižjih nivojih drevesa vozliščem ponavadi ustreza majhno število učnih primerov, zaradi česar se ti preveč prilegajo učni množici in so zato ta vozlišča nezanesljiva. Za rešitev tega se drevesa naknadno poreže. Za rezanje drevesa je v tem delu uporabljena m-ocena.

Pri gradnji odločitvenega drevesa je lahko uporabljena tudi konstruktivna indukcija, kar izboljša rezultate pri odvisnostih atributov.

2.1.3 Naključni gozdovi

Metoda naključnih gozdov je razširitev odločitvenih dreves in poveča točnost napovedi odločitvenega drevesa.

Algoritem zgradi veliko število odločitvenih dreves (100 ali tudi več). Vsako drevo za klasifikacijo novega primera glasuje za svoj razred. Razred, ki dobi največ glasov je izbran.

Gradnja dreves poteka nekoliko drugače kot pri odločitvenih drevesih. Razlika je pri izbiri najboljšega atributa A_i. Namesto, da bi za izbiro najboljšega atributa kandidirali vsi atributi, jih izberemo nekaj. Za število kandidatov se uporablja logaritem števila vseh atributov zvečana za 1, koren števila vseh atributov, ali kar preprosto en sam atribut.

V našem primeru uporabljamo koren števila vseh atributov za določitev kandidatov. Za gradnjo dreves tudi ne uporabljamo naknadnega rezanja dreves in konstruktivne indukcije.

2.1.4 K-najbližjih sosedov

K-najbližjih sosedov(k-NN) je eden preprostejših algoritmov za klasifikacijo in spada v algoritme tako imenovane lene algoritme učenja, ker si le zapomni vse učne primere, vso delo pa prenese na čas klasifikacije.

Postopek klasifikacije poteka tako, da k-NN poišče *k* najbližjih sosedov v atributnem prostoru in nov primer klasificira glede na glasove sosedov.

Pri določanju najbližjih sosedov se za numerične atribute najpogosteje uporablja evklidska razdalja, za diskretne atribute pa zgolj razlika med 0 in 1. Algoritem je možno tudi razširiti, da so glasovi sosedov uteženi glede na njihovo razdaljo od novega primera in z apriorno verjetnostjo razredov.

V našem primeru je uporabljenih 10 sosedov, ki vsi glasujejo z enako močjo.

2.2 Uporabljeni meri za ocenjevanje učenja

2.21 Klasificijska točnost

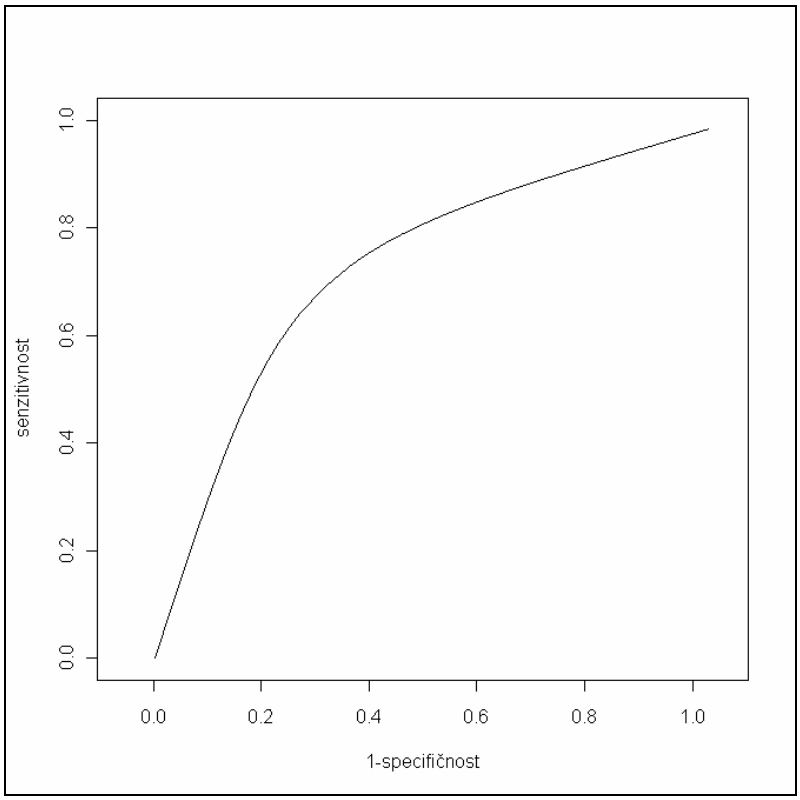
Klasificijsko točnost (angl. classification accuracy) definiramo kot razmerje med vsemi pravilno rešenimi primeri N_p in vsemi primeri N.

$$CA = \frac{N_p}{N} * 100\%$$

Slabost te mere je, da je v primeru, ko je večinski razred zastopan 90%, zelo lahko dobiti trivialni klasifikator, ki ima 90% klasificijsko točnost s klasificiranjem vseh primerov v večinski razred.

2.2.2 Krivulja ROC

Krivulja ROC (ang. Reciever Operating Characteristic), izhaja iz področja teorije odkrivanja signalov in omogoča analizo razmerja med senzitivnostjo in specifičnostjo.
Senzitivnost je definirana kot odstotek pravilno klasificiranih pozitivnih primerov in specifičnost kot odstotek pravilno klasificiranih negativnih primerov.



Slika 2. 1: Primer ROC krivulje

AUC (ang. Area Under the ROC Curve) dobimo tako, da izračunamo površino pod krivuljo ROC, ki je prikazana na sliki 2.1. AUC nam pove, kako dobro klasifikator razloči med pozitivnim in negativnim primerom. Večja kot je vrednost AUC, boljši je klasifikator. Optimalni klasifikator ima AUC oceno 1, naključni pa 0.5, ki ga predstavlja premica iz točke 0.0 do 1.1.

2.3 Friedmanov test

Za primerjavo rezultatov smo izbrali Friedmanov test, ki je opisan v [1]. Osnovna ideja je v tem, da posamezne rezultate rangiramo in nato izračunamo povprečni rang. Friedmanov test testira ničelno hipotezo, ki predpostavlja, da imajo vsi rezultati enak povprečni rang.

Če uspemo zavrniti ničelno hipotezo, lahko nadaljujemo z Nemenyivim testom. S tem testom izračunamo kritično razliko(ang. critical diffrence (CD)), ki nam pove, za koliko se morajo rangi razlikovati, če predpostavimo verjetnost napake prvega reda *p*, da je en rezultat boljši oz. slabši od drugega.

Poglavje 3

3 Obstoječe diskretizacije

Obstoječe metode za diskretizacijo v grobem delimo v dve skupini: nadzorovane in nenadzorovane. Pri nadzorovanih diskretizacija upošteva vrednosti razreda pri nenadzorovani pa ne. Primera nenadzorovanih diskretizacij sta enako-frekvenčna in ekvidistančna. Pod nadzorovane pa sodijo požrešna, Fayyad-Irani, MDL, požrešna z ReliefF-om itd. Diskretizacije z uporabo ReliefF-a lahko poda pravilne meje tudi v primerih močnih odvisnosti med atributi [4].

3.1 Ekvidistančna diskretizacija

Ekvidistančna diskretizacija je ena najpreprostejših diskretizacij, ki pri izbiri mej ne upošteva razreda ter vse meje določi naenkrat.

Metoda deluje tako, da poišče maksimalno in minimalno vrednost atributa ter razpon med njima razdeli na *n* enako širokih intervalov. Pri tem mora biti *n* podan vnaprej kot parameter.

S tem postopkom lahko pade v posamezen interval bistveno več primerov kot v druge. Metoda je tudi zelo občutljiva na skrajne robne primere.

3.2 Enako-frekvenčna diskretizacija

Tudi ta metoda je nenadzorovana in globalna. Ideja te metode je v tem, da imamo intervale z enakim ali skoraj enakim številom vrednosti atributa.

Kot pri ekvidistančni diskretizaciji moramo tudi pri tej vnaprej podati število intervalov *n*. Enako število primerov v vsakem intervalu dosežemo tako, da vrednosti atributa naraščajoče sortiramo in nato v vsak interval po vrsti damo *N / n* vrednosti atributa, kjer je *N* število vseh učnih primerov. Kadar ima atribut veliko število enakih vrednosti, nastane težava saj ni dobro enakih

vrednosti atributa razdeliti v različni diskretni vrednosti. Zaradi tega je treba odstopati od idealne razporeditve.

3.3 Požrešna

Požrešna metoda je za razliko od prejšnjih dveh nadzorovana in lokalna diskretizacija. Obstajata dva pristopa k požrešni diskretizaciji in sicer *od spodaj navzgor* ter *od zgoraj navzdol*.

Pri metodi od spodaj navzgor začnemo z N intervali, kjer N predstavlja število učnih primerov. Te intervale nato združujemo dokler kvaliteta atributa narašča glede na uporabljeno funkcijo merjenja kvalitete atributa. Časovna zahtevnost takega pristopa je $O(N^2)$.

Pri metodi od zgoraj navzdol začnemo z enim samim intervalov in nato iščemo mejo, ki bo največ povečala kvaliteto atributa. Časovna zahtevnost te metode je tudi $O(N^2)$, a v praksi je število dobljenih mej k veliko manjše od števila učnih primerov in se zato časovna zahtevnost zmanjša na $O(kN)$.

Uporabljena požrešna metoda je od spodaj navzgor. Za merjenje kvalitete atributa uporablja informacijski prispevek.

3.4 Fayyad-Irani MDL

Fayyad in Irani sta razvila nadzorovano in lokalno metodo, ki je osnovana na požrešni metodi z nekaj izboljšavami.

Za binarizacijo sta uporabila informacijski prispevek in princip MDL za ustavitev postopka nadaljnje delitve atributov.

Algoritem sta tudi pohitrila z ugotovitvijo, da meja med intervali vedno nastopi med učnimi primeri, ki imajo različen razred. S tem sta metodo pohitrila, ker je teh mej tipično veliko manj od vseh možnih mej.

Uporabljena metoda je narejena po [2] in razširjena tako, da lahko določimo začetne kandidate za meje.

3.5 Požrešna z ReliefF-om

ReliefF je mera za ocenjevanje kvalitete atributa. Od drugih mer se razlikuje po tem, da ne predpostavlja apriorne in pogojne odvisnosti atributov pri danem razredu. Zaradi tega tudi ni kratkovidna. Kot vsako mero za ocenjevanje kvalitete atributov se da tudi to uporabiti za diskretizacijo atributov.

Osnovna ideja algoritma Relief je, da za vsak učni primer poišče najbližji primer iz istega razreda (najbližji zadetek) in najbližji primer iz nasprotnega primer (najbližji pogrešek). Na ta način

oceni kvaliteto atributa glede na lokalne značilnosti razločevanja razredov. Ravno lokalnost pa vključuje v oceno tudi ostale atribute.

Algoritem Relief je naslednji:

Inicializiraj vektor *W* dolžine *a* (število atributov) na 0 ter določi *m* (število iteracij)
For j = 1 **to** m
    Naključno izberi primer *i*
    Primeru *i* poišči najbližji pogrešek *M* in najbližji zadetek *H*
    For att = 1 **to** a
        W[att] <- W[att] - diff(att, i, H)/m + diff(att, i, M)/m
Return W

Kjer je

$$diff(A_i, u_j, u_k) = \begin{cases} \dfrac{\left|v^{(i,j)} - v^{(i,k)}\right|}{Max_i - Min_i}, & A_i \text{ je zveneni} \\ 0, & v^{(i,j)} = v^{(i,k)} \text{ in } A_i \text{ je diskretni}, \\ 1, & v^{(i,j)} \neq v^{(i,k)} \text{ in } A_i \text{ je diskretni} \end{cases}$$

kjer sta u_j in u_k učna primera in $v^{(i,j)}$ vrednost atributa A_i učnega primera u_j.

ReliefF vsebuje tri razširitve:

- deluje pri neznanih vrednostih atributov,
- namesto najbližjega pogreška in zadetka izbere *k* najbližjih pogreškov in zadetkov, kar zmanjša občutljivost na šum v podatkih.
- deluje na večrazrednih problemih.

Požrešna diskretizacija z ReliefF-om je izvedena z algoritmom navedenem v [4].

Poglavje 4

4 Diskretizacija z razvrščanjem

Razvrščanje spada med nenadzorovane metode strojnega učenja. Gre za postopek združevanja učnih primerov, ki so si v nekem smislu podobni[5]. Podobnost primerov je določena glede na razdaljo med učnimi primeri v prostoru atributov. Mere za razdaljo so evklidska razdalja, manhattanska razdalja, Hammingova razdalja itd.

Ideja diskretizacijske metode je v tem, da na učnih primerih naredimo postopek razvrščanja. Ker razvrščanje grupira podobne primere glede na razdaljo med njimi, upamo, da dobljene skupine vsebujejo primere z enakimi razredi. Te skupine, ki vsebujejo podmnožice vseh učnih primerov, uporabimo za določitev začetnih naravnih mej atributov.

Razvrščanje naenkrat upošteva vse atribute, ki jih hočemo diskretizirati, zato pričakujemo, da te meje ohranjajo tudi informacijo o odvisnosti atributov in s tem predlagajo boljše meje od obstoječih metod.

Diskretizacija z razvrščanjem je izpeljana po naslednjem algoritmu:

1. Izbira kandidatov z rojenjem kjer je treba določiti:
    - vrsto rojenja,
    - število skupin,
    - izbira kandidatov za meje,
2. Izbira končnih mej

V nadaljnjih podpoglavjih opišemo nekatere rešitve iz zgornji algoritem. Glede na to, da so vsa vprašanja odprta, se za nadaljevanje uporablja več različic diskretizacije z razvrščanjem.

4.1 Uporabljeni vrsti razvrščanja

Za to nalogo smo preizkušali dve vrsti razvrščanja in sicer k-means razvrščanje in hierarhično razvrščanje.

4.1.1 K-means razvrščanje

K-means razvrščanje skuša dobiti skupine, tako da minimizira vsoto kvadratov znotraj vsake skupine. V splošnem je časovna zahtevnost k-means algoritma naslednja[8]:

- NP poln problem, v *d* dimenzionalnem evklidskem prostoru za dve skupini
- NP poln problem, za poljubno število skupin *k* v 2d ravnini
- $O(n^{dk+1} \log n)$, če sta *k* in *d* podana, kjer je *n* število učnih primerov

Zaradi visoke časovne kompleksnosti optimalne rešitve tega problema se uporablja naslednji algoritem:

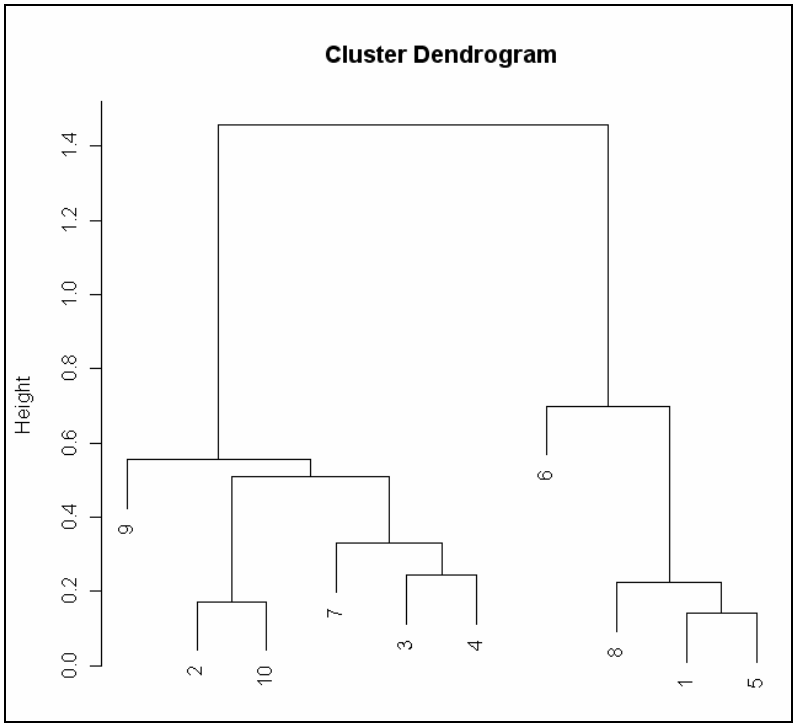
- izberi število skupin *k*,
- naključno generiraj *k* skupin in določi središča le teh, ali kar direktno določi *k* središč,
- vsak učni primer dodaj najbližjemu središču skupine,
- ponovno izračunaj središča novih skupin kot povprečje vseh predstavnikov skupine,
- ponavljaj prejšnja dva koraka dokler središča ne konvergirajo oz. je doseženo maksimalno število iteracij.

S tem algoritmom globalni optimum ni zagotovljen, ker je odvisno od začetnih središč ali bo kongerviral v globalni oz. lokalni optimum. Zaradi te lastnosti, lahko ta algoritem poženemo večkrat z več različnimi začetnimi središči rojev, tako da dobimo bolj zanesljive rezultate. V praksi ima algoritem polinomsko časovno zahtevnost, v najslabšem možnem primeru pa ima eksponentno časovno zahtevnost.

4.1.2 Hierarhično razvrščanje

Hierarhično razvrščanje deluje tako, da za začetek postavi vsak učni primer v svojo skupino, nato pa rekurzivno dve najbližji skupini združi v eno. Dobimo drevo združitev, ki mu pravimo dendrogram, prikazan na sliki 4.1. Iz dendrograma so razvidne vse združitve skupin, pri tem je pomebna tudi višina združitve. Nižja kot je višina združitve prej sta se skupini v postopku združevanja združili. Tako lahko dendrogram enostavno porežemo na določeno višino, da dobimo dano število skupin.

Za praktično reševanje problema najprej sestavimo matriko razdalj, kjer je vrednost v *i*-ti vrstici in *j*-tem stolpcu razdalja med *i*-tim in *j*-tim učnim primerom. Ko rojenje združi dve skupini, se v matriki razdalj to odraža kot združevanje vrstic in stolpcev ter popravkov vrednosti le teh[5].



Slika 4. 1: Primer dendrograma, ki ga vrne hierarhično razvrščanje.

4.2 Določitev števila skupin

Večina postopkov razvrščanja potrebuje kot parameter število skupin, ki ga je v praksi težavno določiti. Število skupin je zelo odvisno od posameznega problema. Uporabili smo prečno preverjanje, tako da smo povečevali število skupin in opazovali klasificijsko natančnost pri napovedovanju z odločitvenimi drevesi. Število skupin smo nehali večati, če je klasificijska točnost padala trikrat zapored. Pri prečnem preverjanju se pri vsakem pregibu izvede izbira končnih mej.

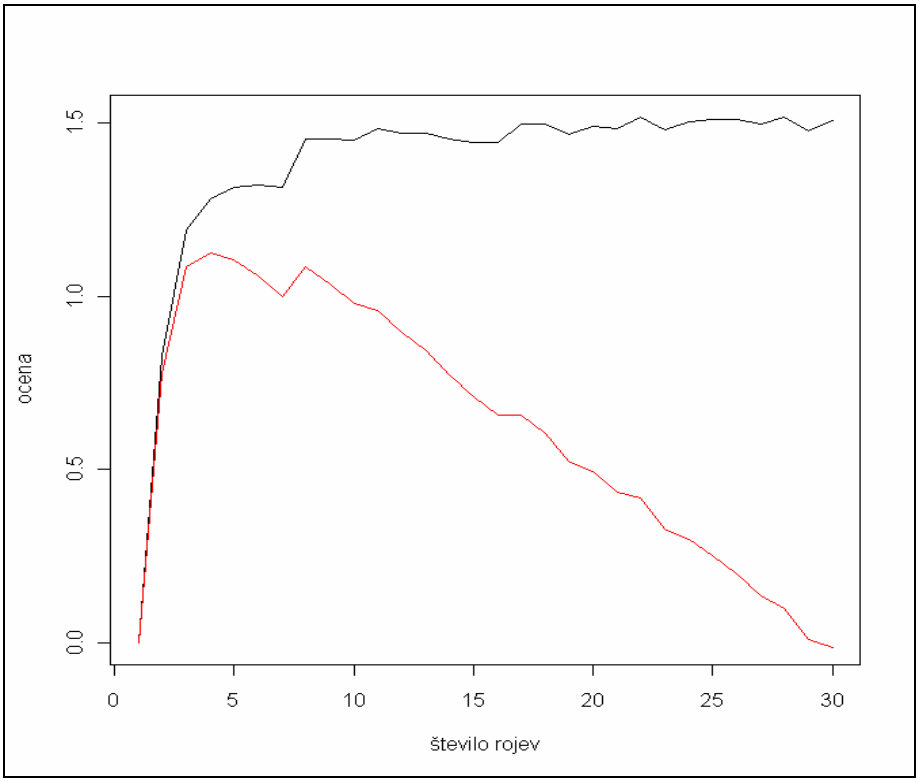
Poizkušali smo tudi z razvrščanjem, ki upošteva razrede, tako da najprej razdelimo učno množico na podmnožice glede na razrede in nato razvrščanje opravimo na teh podmnožicah. Temu nadalje pravimo nadzorovano razvrščanje. V splošnem bi lahko *n* skupin na učni množici s *k* različnimi razredi porazdelili na

$$N = \binom{n-1}{k-1}$$

načinov. To bi prečno preverjanje preveč upočasnilo zato smo vedno uporabili enako število skupin za vsako podmnožico primerov istega razreda.

Število skupin smo poskušali dobiti tudi s hevristiko, ki temelji na prelomnemu(ang. elbow) algoritmu[6]. Za 2 do 30 skupin smo izračunali informacijski in ga utežili glede na število skupin.

To storimo tako, da največji informacijski prispevek razdelimo na 30 delov in nato za vsako skupino odštejemo ta del. Na sliki 4.2 predstavlja črna črta informacijske prispevke glede na število skupin, rdeča pa prikazuje oceno z uteženim informacijskim prispevkom. Iz slike je razvidno, da ocena doseže svoj maksimum pri 4ih skupinah. Ta algoritem je uporaben le za nenadzorovano razvrščanje, saj so pri nadzorovanem razvrščanju skupine vedno čiste in informacijskega prispevka ne moremo izračunati.



Slika 4. 2: Primer ocen pri Elbow hevristiki

4.3 Izbira kandidatov

Ko imamo določene skupine, se moramo odločiti, kako iz njih določiti meje atributov. Ta problem smo rešili tako, da vsaka skupina poda dva kandidata za meje za vsak atribut. Do njih pridemo tako, da elemente skupine sortiramo po danem atributu in maksimalni in minimalni predstavnik predstavlja mejo. Meja je nato še naknadno znižana oz. zvišana do učnega primera, ki pripada drugemu razredu. Po končanem postopku imamo *2ca* mej, kjer je *c* število skupin in *a* število atributov, ki jih diskretiziramo. Veliko teh mej je lahko neuporabnih. Na primer, če imamo atribute, ki so nepomembni še vedno vsaka skupina poda 2 meji za vsak atribut, zato je potrebno te kandidate razredčiti z drugimi algoritmi.

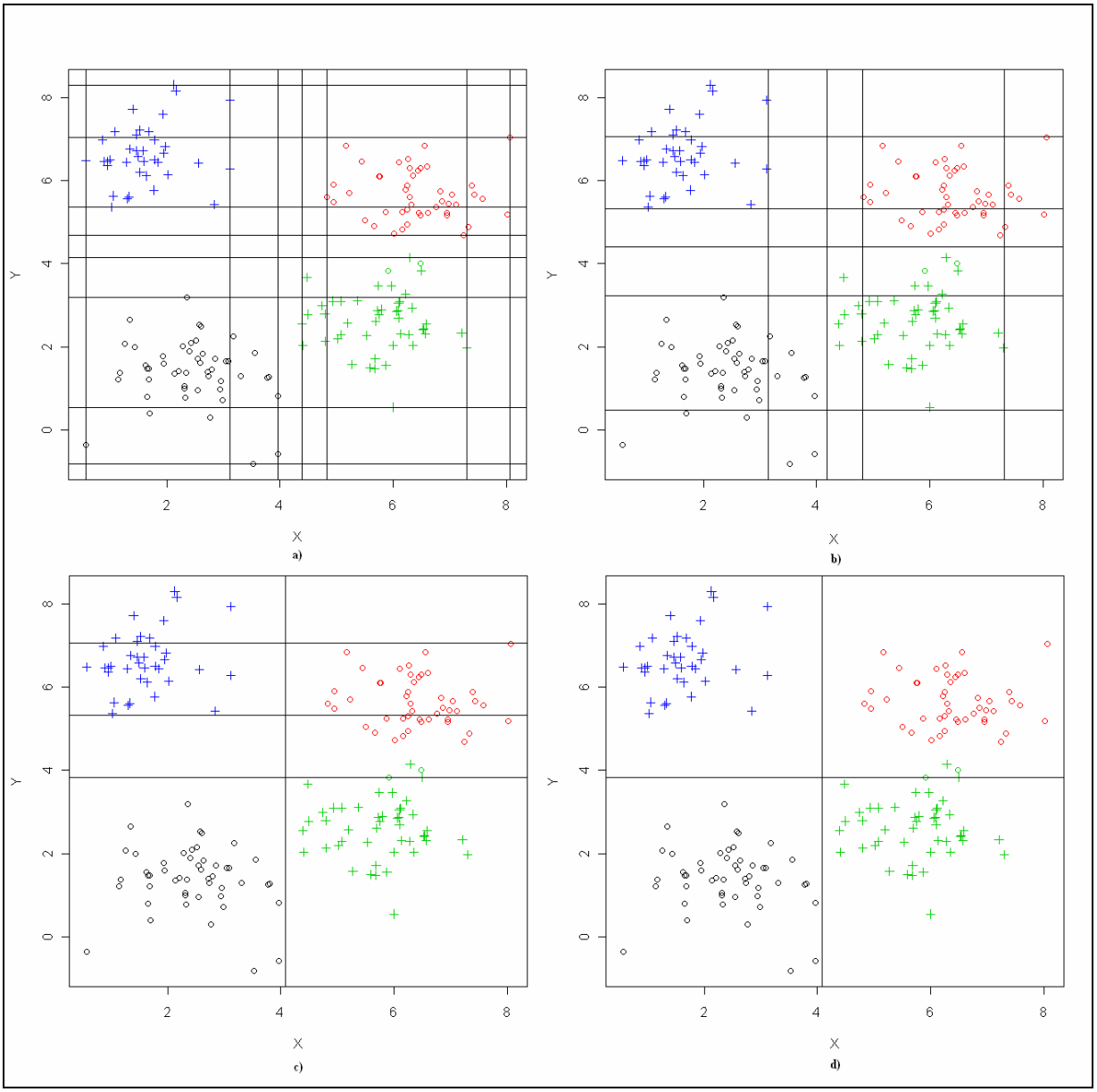
Kandidate za meje naknadno razredčimo tako, da intervale, ki imajo zelo majhno število učnih primerov (npr.: manj kot 5% vseh učnih primerov) odstranimo in njegovi meji nadomestimo s srednjo vrednostjo. V primeru, da ima nek razred zelo malo primerov je bolje, da se temu koraku izognemo, ker lahko s tem odstranimo pomembne kandidate.

Če predpostavimo, da obstaja v podatkih šum, lahko kot parameter določimo tudi procent šuma in tako pri npr. 10% šumu odstranimo 5% maksimalnih in 5% minimalnih primerkov iz skupine in tako kandidate za meje približamo središču skupine.

4.4 Izbira končnih mej

Za izbiro končnih mej smo preizkušali nekaj metod. Prva in najbolj trivialna je kar izbira vseh mej v upanju, da bodo klasifikatorji zaznali dobre meje in izločili neuporabne. Poizkušali smo tudi razredčiti meje z Fayyad-Irani diskretizacijo, požrešno z informacijskim prispevkom in požrešno z ReliefF-om. Poizkušali smo tudi s prečnim preverjanjem kandidatov, tako da smo vedno izbrali kandidata, ki je najbolj povečal klasificijsko točnost klasifikatorja z odločitvenim drevesom.

Na sliki 4.3 vidimo celoten potek izbire mej, kjer barve označujejo različne skupine znaka '+' in 'o' pa predstavljata različne razrede. Slika 4.3a prikazuje meje, ki jih predlagajo skupine, slika 4.3b prikazuje zamik mej do naslednjega primera z različnim razredom, kjer vidimo, da se robne meje odstranijo, ker se zamaknejo v neskončnosti in nekatere notranje meje se zamaknejo na enake vrednosti. Slika 4.3c prikazuje meje, ki ostanejo po združitvi majhnih intervalov in slika 4.4d prikaže končno izbiro mej dobljenih iz kandidatov z ReliefF-om.



Slika 4. 3: Prikaz kandidatov za meje in izbrane končne meje pri uporabi hevristike elbow in izbiranjem končnih mej z ReliefF-om.

Poglavje 5

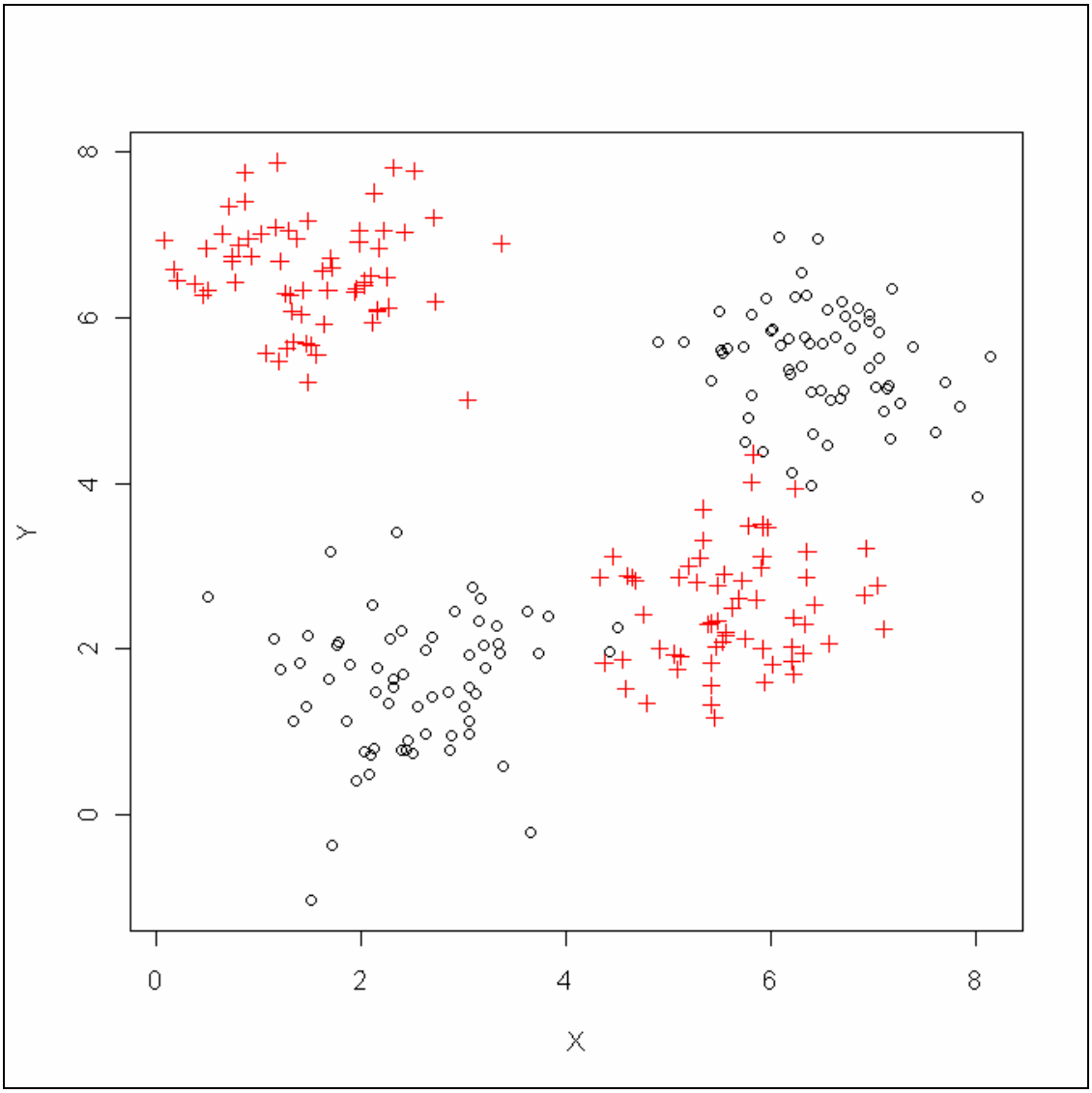
5 Ovrednotenje pristopa

Za izvajanje testov smo uporabili prosto dostopno orodje R. Vsi testi so ponovljeni 10 krat, z razdelitvijo primerov na učno in testno množico v razmerju 70% in 30%. Vsi testi so izpeljani na enakih razdelitvah na učno in testno množico. Opazovali smo klasificijsko točnost in AUC. Zaradi lažje berljivosti so rezultati pomnoženi s 100. Za izdelavo testov so uporabljeni klasifikatorji iz paketa CORElearn.

Najprej smo testirali na treh umetnih množicah z močnimi odvisnostmi med atributi, nato pa še na desetih UCI množicah dobljenih na [9].

5.1 Dvodimenzionalni XOR problem

Da bi videli, kako se diskretizacija z razvrščanjem obnese v primeru, ko imamo opravka z močno odvisnostjo med dvema atributoma, smo ga primerjali na umetni množici z 256 učnimi primeri, od katerih polovica pripada razredu 'o' in druga polovica razredu '+'. Primer je prikazan na sliki 5.1.



Slika 5. 1: Dvodimenzionalni klasificijski primer z močno odvisnostjo med atributoma X in Y

V tabeli 5.1 in 5.2 je z ElbowK-R označena diskretizacija z razvrščanjem, pri kateri je za določitev števila skupin uporabljena hevristika temelječa na Elbow algoritmu, K predstavlja k-means razvrščanje, R označuje ReliefF metodo za izbiro končnih mej iz kandidatov. Podobno velja za ElbowH-R, kjer H predstavlja uporabljeno hierarhično razvrščanje.

Vrsta diskretizacije	Naivni Bayes	Klasificijsko drevo	Naključni gozdovi	k-NN
Ekvidistančna	**78.7**/81.9	87.7/91.8	93.2/98.0	90.4/97.5
Enako frek.	72.9/80.1	83.2/88.5	86.8/94.8	82.9/92.6
Požrešna	70.8/78.6	73.2/73.3	74.6/83.4	76.4/83.7
Fayyad-Irani	73.1/78.7	76.7/78.8	77.8/82.5	75.4/82.3
Vnapr. znanje	45.7/41.4	**99.2**/**99.2**	**99.2**/**99.2**	**99.2**/**99.2**
ReliefF	78.2/**86.7**	86.0/89.7	88.7/95.4	84.8/93.7
ElbowK-R	61.8/60.6	89.5/91.4	89.7/92.9	90.2/93.0
ElbowH-R	64.1/64.0	92.3/94.2	92.8/95.6	91.8/95.5

Tabela 5. 1: Klasificijska točnost in AUC za različne diskretizacije in Elbow hevristiko.

	Št. skupin/Št. mej
ElbowK-R	4.1/4.6
ElbowH-R	4.2/4.8

Tabela 5. 2: Število skupin in dobljenih končnih mej pri Elbow hevristiki.

V vrstici vnaprejšnje znanje sta pri obeh atributih uporabljeni meji na vrednosti 4. Tabela 5.2 prikazuje povprečno število skupin, ki ga predlaga Elbow hevristika, ter število končnih mej.

V tabelah 5.3, 5.4, 5.5, 5,6 so testi izvedeni za diskretizacijo z razvrščanjem, pri čemer je število skupin določeno s prečnim preverjanjem, uporabljeni pa so različni algoritmi za izbiro končnih mej.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	64.3/63.6	94.3/96.4	95.1/98.3	94.4/98.2	4.0/5.1
Požrešna	65.6/62.4	**95.7**/**97.1**	**96.1**/**98.5**	**96.1**/**98.6**	4.2/5.0
Fayyad-Irani	65.7/66.5	66.2/65.5	66.2/67.7	64.8/66.8	6.6/1.6
CV	43.5/51.4	94.4/97.0	94.5/97.4	94.7/97.0	5.4/2.7
ReliefF	**66.4**/**69.5**	91.9/94.5	93.5/96.5	90.9/96.7	4.3/5.0

Tabela 5. 3: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	**67.7**/66.0	94.4/ 97.1	95.3/**99.1**	95.2/**99.1**	4.1/4.8
Požrešna	65.7/ 65.1	94.9/97.4	95.6/98.9	95.8/98.8	3.9/4.5
Fayyad-Irani	63.1/66.4	62.9/64.7	63.9/67.6	64.0/67.5	6.9/2.1
CV	48.8/49.2	**96.9**/**97.9**	**97.0**/98.2	**95.6**/98.0	5.0/2.3
ReliefF	66.5/**74.8**	93.0/94.6	93.6/97.8	91.4/97.9	4.7/5.8

Tabela 5. 4: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s hierarhičnim razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	59.6/61.0	96.9/98.0	97.7/99.2	97.7/99.3	4.0/4.0
Požrešna	59.6/61.0	96.9/98.0	97.7/99.2	97.7/99.3	4.0/4.0
Fayyad-Irani	**69.0**/**71.6**	70.9/72.4	71.6/75.0	69.6/73.5	10.2/2.5
CV	44.0/42.4	**99.1**/**99.2**	**99.1**/**99.2**	**99.1**/**99.1**	4.2/2.0
ReliefF	60.1/67.1	88.1/90.4	89.9/94.5	87.8/93.9	5.6/3.0

Tabela 5. 5: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	59.6/61.0	96.9/98.0	97.7/99.2	97.7/**99.3**	4.0/4.0
Požrešna	59.6/61.0	96.9/98.0	97.7/99.1	97.7/**99.3**	4.0/4.0
Fayyad-Irani	72.2/76.1	73.0/76.4	73.5/79.6	71.2/77.0	9.6/3.3
CV	48.7/44.9	**98.6**/**98.9**	**98.7**/**99.3**	**98.7**/99.1	4.8/2.3
ReliefF	**74.3**/**78.2**	88.8/91.3	90.8/94.6	89.7/94.5	6.2/4.2

Tabela 5. 6: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim hierarhičnim razvrščanjem.

Kot vidimo iz tabel 5.1, 5.3, 5.4, 5.5 in 5.6 se naivni Bayesov klasifikator najslabše izkaže, kar je pričakovano zaradi njegove naivnosti. Vsi drugi uporabljeni klasifikatorji znajo pravilne meje uporabiti za izboljšanje napovedovanja.

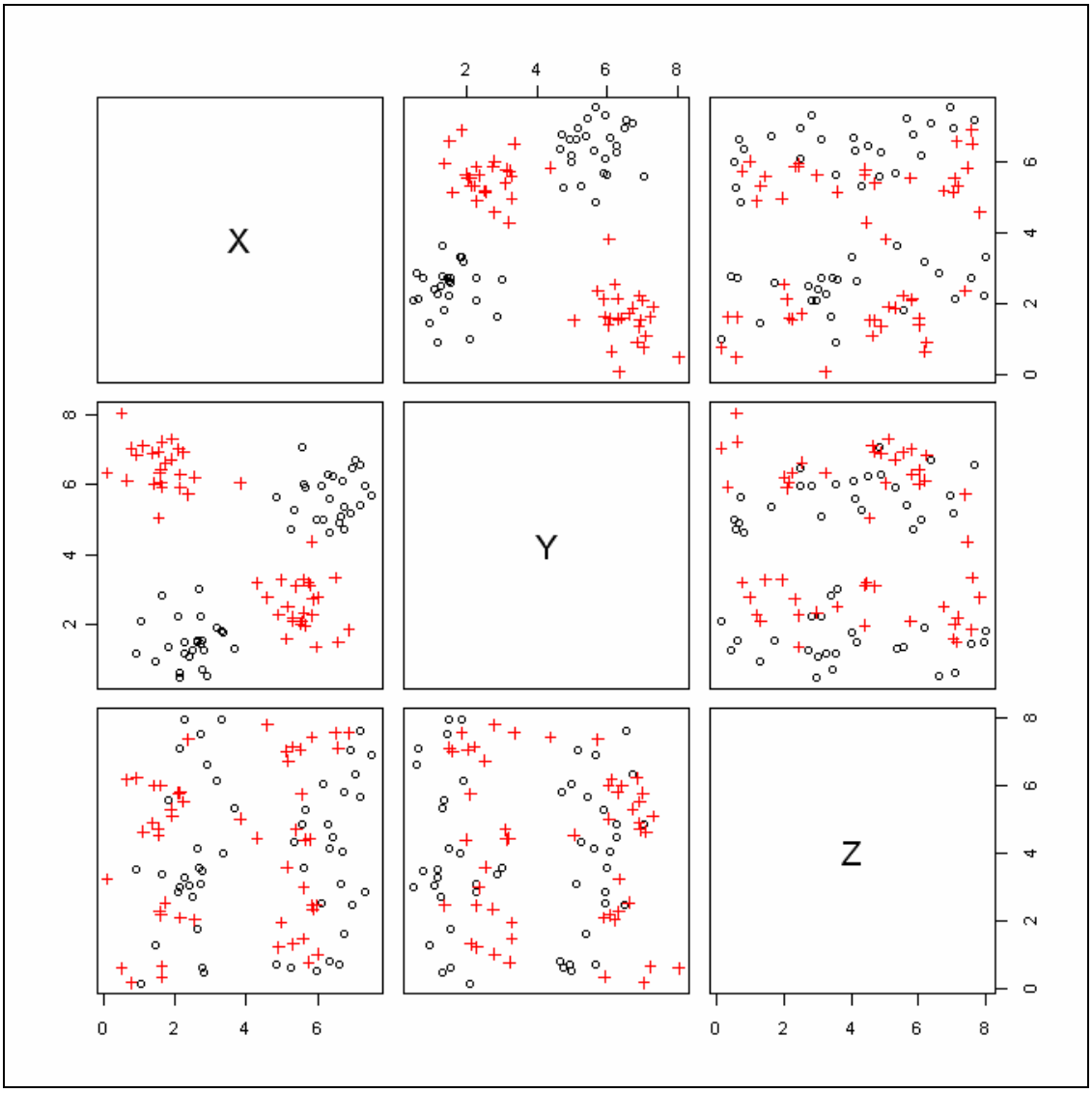
Iz teh tabel 5.1 – 5.6 je tudi razvidno, da se skoraj pri vseh metodah z razvrščanjem izberejo boljše meje od ostalih diskretizacij. Slabo se obnesejo le pri uporabi algoritma Fayyad-Irani za določitev končnih mej. Čeprav Fayyad-Irani izbere majhno število mej, pa zaradi svoje kratkovidnosti te niso informativne za problem z močnimi odvisnostmi. V povprečju dobimo s Fayyad-Irani metodo za izbiro končnih mej več skupin kot pri ostalih metodah.

Pri požrešni metodi in metodi z ReliefF-om vidimo, da s predhodnim razvrščanjem izboljšamo rezultate, kar da slutiti, da razvrščanje v tem primeru odpravi nekatere nepravilne kandidate za meje.

Nekoliko boljše rezultate je opaziti pri uporabi nadzorovanega razvrščanja, razlike med uporabo različnih vrst razvrščanja pa so neopazne. Najbolje se v vseh primerih izkaže uporaba prečnega preverjanja za določitev števila rojev ter prečno preverjanje za izbiro končnih mej, ki pa je tudi najbolj časovno zahtevna metoda med vsemi.

5.2 Dvodimenzionalni XOR problem z neuporabnim atributom

Problem je podoben prejšnjemu, le da smo dodali še dodatni atribut Z, ki je naključno porazdeljen in ne vsebuje nobene informacije. Ker razvrščanje poda enako število kandidatov za meje za vse atribute, hočemo videti katera metoda za izbiro končnih mej je sposobna odpraviti neuporabne meje.



Slika 5. 2: 2D primer z močno odvisnostjo med atributoma X in Y ter neuporabnim atributom Z.

Za vnaprejšnje znanje sta uporabljeni meji na vrednosti 4 za atributa X in Y, atribut Z pa nima mej.

Vrsta diskretizacije	Naivni Bayes	Klasificijsko drevo	Naključni gozdovi	k-NN
Ekvidistančna	62.9/67.7	75.3/82.1	80.7/88.7	71.1/81.7
Enako frek.	55.3/60.2	62.2/63.1	63.0/71.5	60.7/67.8
Požrešna	56.356.1	58.4/59.4	59.8/67.1	55.9/61.3
Fayyad-Irani	57.3/55.6	56.9/55.2	57.3/55.5	56.4/55.6
Vnapr. znanje	37.6/39.0	**100/100**	**100/100**	**100/100**
ReliefF	65.3/72.6	68.7/73.4	71.3/78.8	68.0/74.6
ElbowK-R	**66.1/73.3**	72.1/75.0	74.2/81.4	65.6/70.6
ElbowH-R	65.2/72.1	71.2/75.1	73.8/80.7	63.4/69.7

Tabela 5. 7: Klasificijska točnost in AUC za različne diskretizacije in Elbow hevristiko

	Št. skupin/Št. mej
ElbowK-R	8.3/15.7
ElbowH-R	8.3/16.8

Tabela 5. 8: Število skupin in dobljenih končnih mej pri Elbow hevristiki.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	52.7/59.3	72.7/75.8	73.3/78.9	73.3/78.6	4.5/11.6
Požrešna	61.0/65.7	77.7/80.0	78.0/81.7	72.0/78.7	4.6/11.7
Fayyad-Irani	58.3/53.8	57.7/53.2	58.3/53.8	57.7/53.8	3.2/0.7
CV	59.7/56.6	**80.3/84.6**	**82.3/87.8**	**81.7/87.6**	3.6/3.0
ReliefF	**63.7/68.0**	73.0/74.5	75.3/82.4	72.7/75.8	5.1/11.1

Tabela 5. 9: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	60.0/64.2	69.3/74.8	73.3/79.5	68.3/75.4	4.0/10.0
Požrešna	**62.7/68.7**	76.0/81.1	79.0/83.3	72.3/79.2	5.3/13.4
Fayyad-Irani	56.7/52.4	56.7/52.4	56.7/52.4	56.7/52.4	2.8/0.5
CV	47.3/50.1	**88.0/92.9**	**88.7/93.3**	**86.7/92.2**	5.0/9.2
ReliefF	58.0/64.5	77.3/81.0	78.8/86.2	73.3/83.4	4.1/7.5

Tabela 5. 10: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s hierarhičnim razvrščanju.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	**66.3/71.6**	82.7/84.8	84.7/85.3	78.0/85.7	4.8/11.4
Požrešna	62.3/66.5	83.7/84.9	80.7/83.5	76.0/83.4	5.4/12
Fayyad-Irani	59.3/55.2	59.3/55.2	59.3/55.2	59.3/55.2	4.4/0.8
CV	59.7/56.5	**89.7/91.6**	**90.0/94.1**	**86.7/93.0**	7.4/3.8
ReliefF	61.3/70.2	81.7/82.7	83.7/83.9	78.0/86.0	5.0/10

Tabela 5. 11: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	62.0/68.8	84.3/84.7	86.3/86.4	80.7/87.1	4.2/10.3
Požrešna	63.7/69.5	81.7/82.0	82.7/85.7	76.0/84.2	5.0/11.5
Fayyad-Irani	59.7/55.5	59.7/55.5	59.7/55.5	59.7/55.5	4.8/0.7
CV	52.3/56.9	**91.7/94.1**	**92.0/95.7**	**89.7/94.9**	7.0/4.0
ReliefF	**65.7/71.7**	81.3/82.6	84.3/85.6	81.7/87.8	5.8/9.8

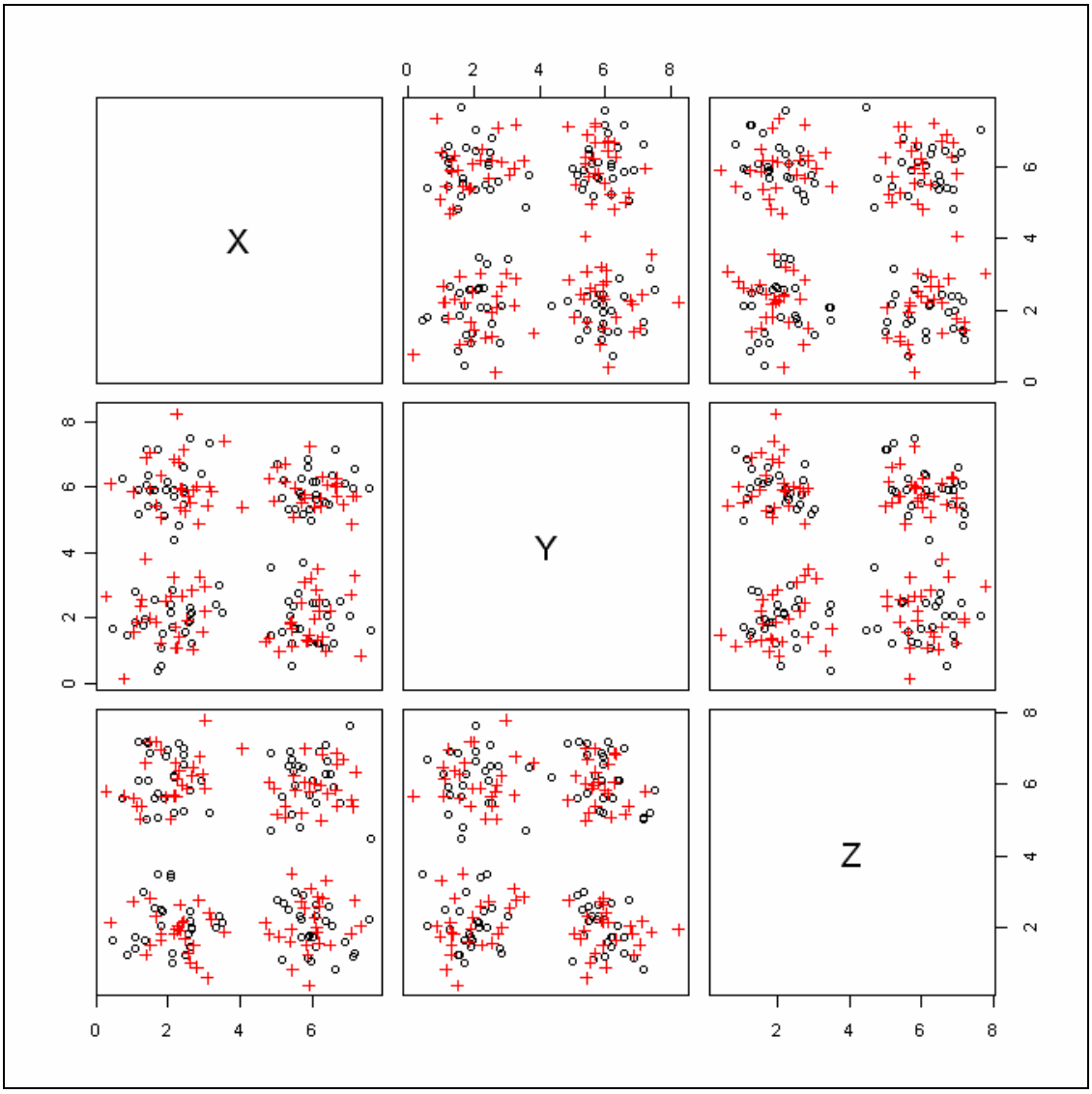
Tabela 5. 12: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim hierarhičnim razvrščanjem.

Iz tabel 5.7 ter 5.9-5.12 je opaziti, da so vse metode občutljive na neuporaben atribut, čeprav lahko z vnaprejšnjem ekspertnim znanjem dosežemo na tem umetnem primeru 100 klasificijsko točnost in AUC(razen za naivnega Bayesa).

Na tem primeru lahko vidimo, da se pri izbiri končnih mej iz kandidatov pri diskretizaciji z razvrščanjem metoda Fayyad-Irani spet najslabše izkaže in v povprečju izbere manj kot 1 mejo, kar pomeni, da zavrne preveč kandidatov. Nasprotno pa požrešna metoda ne zavrne skoraj nobenega kandidata. Spet se za najboljšo metodo za izbiro končnih mej izkaže prečno preverjanje.

5.3 Tridimenzionalni XOR problem

Za zadnji umetni test smo izbrali primer močne odvisnosti med tremi atributi kot prikazuje slika 5.3. Iz slike vidimo, da z dvema atributoma ne moremo napovedati ničesar.



Slika 5. 3: 3D XOR primer z močnimi odvisnostmi med atributi X, Y in Z.

Vrsta diskretizacije	Naivni Bayes	Klasificijsko drevo	Naključni gozdovi	k-NN
Ekvidistančna	48.6/51.4	65.6/70.7	71.8/79.0	49.5/48.1
Enako frek.	49.6/51.4	51.9/53.3	55.3/56.7	44.8/42.9
Požrešna	50.4/50.2	49.2/49.4	48.4/51.2	53.0/53.7
Fayyad-Irani	49.0/50.0	49.0/50.0	49.0/50.0	49.0/50.0
Vnapr. znanje	42.2/40.6	**98.6/98.4**	**98.6/98.4**	**98.6/98.6**
ReliefF	**56.2/57.7**	57.7/59.8	57.5/60.4	52.3/50.8
ElbowK-R	45.8/46.0	91.1/93.0	92.2/95.9	89.6/93.9
ElbowH-R	44.3/45.9	93.1/96.3	92.9/97.1	89.8/95.4

Tabela 5. 13: Klasificijska točnost in AUC za različne diskretizacije in Elbow hevristiko.

Iz tabele 5.13 vidimo, da se vse preverjene obstoječe diskretizacije slabo izkažejo glede na vnaprejšnje znanje. Metodi z razvrščanjem predlagata boljšo diskretizacijo, kar izkoristiti vsi algoritmi razen naivnega Bayesa.. Vidimo, da se vse metode v tabelah 5.15, 5.16, 5.17 ter 5.18 dobro odrežejo, razen, ko končne meje izbiramo z metodo Fayyad-Irani, ki zavrne vse kandidate.

Različica s prečnim preverjanjem za določbo števila skupin in prečnim preverjanjem za izbiro končnih mej deluje tukaj skoraj idealno saj v povprečju izbere le 3.2 meje, kar je zelo blizu vnaprejšnjemu znanju, ki vsebuje 3 meje in sicer za vsak atribut pri vrednosti 4.

	Št. skupin/Št. mej
ElbowK-R	8.4/4.7
ElbowH-R	8.4/4.8

Tabela 5. 14: Število rojev in dobljenih končnih mej pri Elbow hevristiki.

Uporabljena metoda za izbiro končnih mej	**Naivni Bayes**	**Odločitveno drevo**	**Naključni gozdovi**	**k-NN**	**Št. skupin/ Št. mej**
Brez	48.3/46.9	90.7/93.8	92.2/96.2	85.8/92.6	5.8/5.7
Požrešna	**49.5**/47.4	92.3/94.9	92.3/96.1	89.7/94.1	5.4/5.0
Fayyad-Irani	48.7/**50.0**	48.7/50.0	48.7/50.0	48.7/50.0	2.0/0.0
CV	44.2/43.0	**97.8**/**98.3**	**97.8**/**98.3**	**97.8**/**98.4**	6.7/3.2
ReliefF	48.3/46.6	92.5/95.0	94.8/95.5	91.0/95.5	6.1/4.4

Tabela 5. 15: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	**Naivni Bayes**	**Odločitveno drevo**	**Naključni gozdovi**	**k-NN**	**Št. skupin/ Št. mej**
Brez	47.8/45.6	84.3/87.8	86.8/91.5	85.5/90.5	5.3/4.9
Požrešna	48.7/46.4	91.0/93.3	93.5/97.6	91.2/96.0	5.1/4.8
Fayyad-Irani	48.7/**50.0**	48.7/50.0	48.7/50.0	48.7/50.0	2.0/0.0
CV	42.7/42.9	**96.8**/**97.9**	**97.0**/**98.0**	**97.5**/**97.6**	6.3/3.2
ReliefF	48.3/45.5	89.8/92.6	92.3/97.4	90.5/95.3	5.6/5.0

Tabela 5. 16: Klasificijske točnosti in AUC pri izbiri števila skupin s CV s hierarhičnim razvrščanjem.

Uporabljena metoda za izbiro končnih mej	**Naivni Bayes**	**Odločitveno drevo**	**Naključni gozdovi**	**k-NN**	**Št. skupin/ Št. mej**
Brez	44.5/43.8	88.2/89.7	89.3/91.6	89.0/90.6	5.4/4.0
Požrešna	44.3/45.0	94.8/97.3	96.0/99.0	94.2/98.3	6.0/4.0
Fayyad-Irani	48.7/**50.0**	48.7/50.0	48.7/50.0	48.7/50.0	4.6/0.0
CV	43.8/42.1	**97.2**/**98.1**	**97.5**/**98.3**	**97.5**/**97.8**	7.0/3.2
ReliefF	47.3/47.6	96.2/96.7	95.2/97.3	94.8/96.9	6.0/3.7

Tabela 5. 17: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim k-means razvrščanjem.

Uporabljena metoda za izbiro končnih mej	Naivni Bayes	Odločitveno drevo	Naključni gozdovi	k-NN	Št. skupin/ Št. mej
Brez	43.7/42.5	96.0/96.8	95.7/97.4	95.3/96.9	6.6/3.8
Požrešna	44.5/43.1	96.0/96.9	95.7/97.7	95.3/97.3	6.6/3.8
Fayyad-Irani	48.7/**50.0**	48.7/50.0	48.7/50.0	48.7/50.0	4/0.0
CV	43.7/42.6	**97.7**/**98.0**	**98.0**/**98.5**	**98.2**/**98.4**	6.8/3.1
ReliefF	44.2/41.5	96.2/96.7	96.3/97.7	96.0/97.4	6.4/3.5

Tabela 5. 18: Klasificijske točnosti in AUC pri izbiri števila skupin s CV z nadzorovanim hierarhičnim razvrščanjem.

5.4 Primeri iz UCI

Za testiranje smo uporabili 10 množic iz tabele 5.19. Vse množice razen Adult so uporabljene v celoti, pri tej pa je uporabljena naključno izbrana množica 500ih učnih primerov.

Domena	Število atributov	Št. zveznih atributov	Št. vrednosti razreda	Večinski razred	Število primerov
Adult	15	6	2	76.2	10971
Annealing	39	6	5	76.2	798
Breast tissue	10	9	6	20.8	106
Credit app.	16	6	2	55.1	666
Glass	10	9	6	35.5	214
Heart cleve.	14	5	5	54.1	303
Ionosphere	35	34	2	64.1	351
Iris	5	4	3	33.3	150
Statlog(Heart)	14	6	2	55.6	270
Wine	14	13	3	39.9	178

Tabela 5. 19: Podatki o uporabljenih domenah iz UCI.

V tabelah 5.20 do 5.27 so prikazane klasificijske točnosti in pripadajoče standardni odkloni(SD) za dane domene, v tabelah 5.28 do 5.35 pa je prikazan AUC in standardne deviacije.

Vse diskretizacije z razvrščanjem v teh tabelah vsebujejo za določanje števila skupin prečno preverjanje razen različici z oznako EK in EH. Ti dve uporabljata Elbow hevristiko. Za izbiranje končnih mej je pri vseh razen pri Kcv in Hcv uporabljen ReliefF, pri teh dveh pa se končne meje določijo z prečnim preverjanjem. Ks in Hs za razliko od ostalih uporabljata nadzorovano razvrščanje.

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	79.9/1.8	79.2/3.1	77.6/2.8	**80.0/2.8**	79.3/2.7	76.0/3.5	77.7/3.4
Annealing	**81.0/3.8**	74.5/5.2	79.2/3.2	77.1/4.9	76.6/3.9	77.5/3.1	77.9/4.7
Breast tissue	53.8/7.6	62.5/5.7	60.0/9.5	58.1/9.1	**65.0/7.9**	61.6/5.9	55.3/9.0
Credit app.	83.9/1.7	84.5/2.0	83.8/2.1	84.5/1.5	**84.9/2.0**	83.3/1.1	81.3/2.6
Glass	52.0/7.7	52.6/10.9	56.0/6.9	58.0/6.9	**60.2/10.3**	50.6/8.3	50.9/10.9
Heart cleve.	55.8/3.4	55.9/4.6	51.1/5.4	56.0/4.0	55.2/2.9	**56.5/4.1**	56.0/4.8
Ionosphere	88.5/1.8	87.3/1.7	79.2/4.3	88.2/2.7	88.4/2.9	86.9/4.3	88.5/2.0
Iris	96.0/2.5	90.0/4.9	95.3/2.4	95.3/3.0	94.7/3.2	95.3/3.4	96.0/2.0
Statlog(Heart)	**83.1/3.5**	82.8/3.0	81.1/4.1	81.7/2.9	82.3/3.0	78.0/4.2	81.0/2.3
Wine	90.7/4.9	90.7/4.3	87.0/4.2	89.6/5.9	91.1/3.4	85.2/7.2	90.4/4.5

Tabela 5. 20: Klasificijska točnost in pripadajoči SD za naivni Bayesov klasifikator. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	77.9/3.1	78.5/3.4	76.3/2.6	76.9/3.3	79.1/3.1	78.9/3.2
Annealing	74.3/3.6	75.4/5.1	42.2/43.45	73.2/25.1	74.7/4.3	74.9/3.6
Breast tissue	59.7/8.7	59.4/7.2	53.1/12.6	49.1/8.2	62.5/8.1	56.2/10.0
Credit app.	84.4/1.2	84.3/2.1	79.5/9.9	83.8/1.7	83.7/1.8	83.9/1.7
Glass	59.1/10.6	56.2/10.0	47.8/7.6	52.2/11.6	54.8/6.2	53.7/12.4
Heart cleve.	55.2/2.7	55.5/5.2	54.3/4.1	55.8/4.8	55.9/4.4	**56.5/3.8**
Ionosphere	88.6/2.8	**97.0/2.8**	80.5/8.0	84.4/7.1	87.7/3.2	83.2/3.9
Iris	96.4/3.0	96.0/3.9	78.7/11.4	66.4/11.5	96.7/2.4	**97.1/3.0**
Statlog(Heart)	80.9/4.0	81.5/3.6	72.1/6.4	77.4/3.0	81.4/3.8	81.1/2.7
Wine	**91.9/3.8**	89.4/6.6	85.0/10.7	84.1/7.4	90.0/4.2	90.2/4.5

Tabela 5. 21: Klasificijska točnost in pripadajoči SD za naivni Bayesov klasifikator. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	74.4/4.5	75.1/4.6	59.4/1.7	**77.1/4.5**	75.9/6.0	74.8/4.3	76.4/2.5
Annealing	93.6/1.7	92.3/1.8	**96.7/1.4**	96.2/1.1	**96.7/1.2**	92.8/2.3	92.4/1.7
Breast tissue	49.7/13.7	55.3/9.1	55.2/7.5	57.8/7.1	55.9/9.3	51.6/9.8	56.6/8.5
Credit app.	84.1/2.1	84.7/1.6	84.1/1.5	84.2/2.6	**85.2/2.1**	84.2/1.9	84.0/2.0
Glass	49.5/4.9	59.7/9.3	52.5/10.0	**60.3/6.6**	54.9/10.1	55.7/8.2	56.5/7.0
Heart cleve.	52.1/4.1	51.9/3.9	51.1/5.8	48.5/5.7	49.0/6.9	**52.4/3.2**	48.7/6.8
Ionosphere	86.2/2.0	89.2/4.1	80.7/4.2	87.5/3.0	88.1/3.4	87.7/2.7	88.1/3.3
Iris	96.0/2.9	90.0/5.6	95.3/2.2	95.6/1.8	96.2/2.1	95.6/3.9	**97.6/1.3**
Statlog(Heart)	78.1/4.2	78.1/4.7	77.8/4.7	78.5/2.8	78.1/4.4	73.8/4.4	76.9/3.8
Wine	89.4/5.0	92.4/4.2	75.9/9.2	92.8/3.5	92.4/4.8	91.5/4.5	88.9/7.2

Tabela 5. 22: Klasificijska točnost in pripadajoči SD za odločitvena drevesa. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	75.1/3.3	76.5/3.1	76.9/2.3	76.3/2.9	75.6/5.3	76.1/4.2
Annealing	92.5/2.4	92.5/1.8	47.0/48.47	83.9/28.7	93.6/2.5	93.1/2.7
Breast tissue	52.5/13.3	57.2/10.8	45.1/10.9	49.4/11.0	56.2/11.6	**58.1/12.2**
Credit app.	84.7/2.2	84.2/2.8	80.5/10.2	84.1/1.9	85.0/1.4	84.0/2.0
Glass	55.1/10.0	52.0/8.2	47.2/8.0	53.4/9.3	52.6/8.0	58.6/8.0
Heart cleve.	50.2/6.6	49.8/4.5	50.1/3.9	50.4/3.5	51.9/4.6	50.1/4.5
Ionosphere	**89.6/2.9**	89.3/3.1	81.2/10.3	86.9/3.7	88.2/1.7	87.2/3.1
Iris	96.9/1.9	96.0/2.3	79.6/12.2	68.2/12.7	97.1/1.8	97.3/1.4
Statlog(Heart)	75.9/5.6	77.2/5.1	73.2/6.5	**79.9/3.0**	73.7/5.4	76.4/4.3
Wine	92.0/2.9	**93.5/3.5**	88.0/6.0	88.5/7.7	89.6/5.5	88.0/4.2

Tabela 5. 23: Klasicijska točnost in pripadajoči SD za odločitvena drevesa. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	79.3/3.1	80.3/2.9	72.9/3.7	**80.5/2.8**	80.0/3.6	77.1/2.4	78.7/3.3
Annealing	95.0/2.3	93.1/1.3	**97.2/1.1**	96.8/1.5	97.0/1.1	93.0/1.5	92.4/1.4
Breast tissue	61.3/10.9	63.7/6.5	57.2/5.7	**66.2/6.2**	63.1/8.2	61.3/8.5	63.1/10.6
Credit app.	84.5/1.8	85.8/1.9	**86.5/1.5**	85.1/1.9	86.1/1.9	85.7/1.8	84.9/2.3
Glass	56.5/8.9	70.3/5.6	64.3/5.7	62.9/5.3	**71.1/5.0**	60.5/8.7	62.0/8.5
Heart cleve.	58.0/3.6	**58.2/4.0**	56.5/3.4	56.7/3.6	57.1/3.4	57.5/4.0	57.5/3.5
Ionosphere	**92.5/2.0**	91.2/2.7	86.9/2.9	91.4/2.3	91.1/1.9	90.6/3.1	91.2/2.1
Iris	95.8/2.9	93.3/3.5	96.7/1.9	95.1/2.7	95.6/3.1	95.8/3.4	96.7/1.9
Statlog(Heart)	**84.0/2.6**	82.5/2.6	79.8/4.5	82.1/2.7	82.2/4.0	77.4/4.1	81.0/3.7
Wine	**98.0/2.4**	95.9/2.4	93.5/4.0	96.1/3.7	95.7/3.7	94.6/2.8	94.6/5.1

Tabela 5. 24: Klasicijska točnost in pripadajoči SD za naključne gozdove. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	78.5/3.2	79.0/3.5	77.2/3.9	77.7/3.2	79.5/4.1	78.8/4.3
Annealing	94.1/2.0	94.7/1.9	47.8/49.26	84.0/28.7	93.4/2.0	92.8/1.9
Breast tissue	56.2/12.4	58.4/9.2	55.3/10.1	55.0/9.7	59.4/6.8	63.7/10.5
Credit app.	85.2/1.5	85.5/2.4	81.7/10.6	84.9/1.5	85.1 /.8	85.8/1.9
Glass	68.2/8.8	66.5/7.5	49.1/9.4	54.6/11.3	66.6/8.4	64.3/11.6
Heart cleve.	56.9/4.3	57.0/3.6	55.9/4.9	57.4/4.3	56.8/3.1	57.6/2.8
Ionosphere	91.0/2.4	91.1/2.0	82.0/9.2	87.7/3.6	90.8/2.1	89.4/3.1
Iris	96.9/1.9	96.2/2.1	80.0/12.4	68.2/12.7	**97.1/1.8**	96.9/2.6
Statlog(Heart)	81.5/4.0	81.5/3.6	75.2/7.3	80.0/4.2	79.9/3.9	81.1/3.9
Wine	96.3/2.8	96.1/3.2	90.6/6.1	88.9/7.7	95.6/2.7	96.9/3.2

Tabela 5. 25: Klasicijska točnost in pripadajoči SD za naključne gozdove. (2. del)

Domena	Ekvidist.	En. frek.	Pozrešna	Fay.-Ir.	ReliefF	K	H
Adult	79.3/3.0	79.8/3.9	78.7/3.7	**80.2/2.1**	77.8/4.0	75.5/2.3	77.6/4.5
Annealing	94.9/1.4	93.8/1.6	**96.4/1.7**	95.8/1.7	95.8/1.3	93.4/2.3	93.3/1.7
Breast tissue	54.1/9.7	56.2/8.1	54.1/8.3	**60.3/9.2**	**60.3/7.8**	58.4/7.2	55.6/10.3
Credit app.	83.7/2.1	**86.2/1.8**	**86.2/1.8**	84.4/2.3	85.2/2.1	83.9/1.2	83.9/2.0
Glass	50.8/7.9	62.8/7.1	63.5/8.4	63.7/5.2	63.8/5.4	58.3/7.7	57.1/8.0
Heart cleve.	56.9/4.8	56.6/4.6	55.7/4.1	55.5/2.3	56.4/3.3	57.0/3.0	57.3/3.1
Ionosphere	**90.7/2.1**	90.3/1.7	87.1/3.1	88.0/2.3	89.7/2.6	88.1/2.5	88.5/3.8
Iris	92.9/2.5	92.7/4.7	94.2/2.4	96.2/3.3	95.1/3.9	94.4/5.3	**97.1/2.1**
Statlog(Heart)	80.4/3.3	80.6/3.9	79.0/2.9	80.9/3.6	80.1/2.6	77.4/3.5	79.4/2.7
Wine	95.0/2.8	96.5/2.5	**97.4/2.0**	94.6/3.9	95.9/4.2	92.8/4.8	91.9/5.1

Tabela 5. 26: Klasificijska točnost in pripadajoči SD za k-NN klasifikator. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	78.1/4.2	78.0/3.0	75.9/2.8	76.9/2.6	79.0/3.8	78.0/3.8
Annealing	93.9/1.5	94.2/1.4	47.7/49.16	84.4/28.8	94.0/2.1	93.6/1.7
Breast tissue	51.2/9.6	51.9/6.9	54.4/10.9	49.7/9.5	55.9/8.3	53.4/8.5
Credit app.	84.3/2.3	85.8/1.8	80.9/10.2	84.6/1.2	84.7/2.2	84.2/1.5
Glass	**66.8/5.9**	66.2/8.9	49.1/9.2	56.3/6.3	61.8/6.0	59.2/8.8
Heart cleve.	55.5/3.7	57.4/3.8	55.3/4.3	55.6/2.9	57.1/3.6	**57.9/3.1**
Ionosphere	89.4/2.1	88.4/4.1	80.4/9.2	86.5/4.2	88.3/5.0	86.8/4.4
Iris	96.2/2.4	96.7/2.6	75.8/17.4	63.1/15.2	95.1/5.6	**97.1/2.1**
Statlog(Heart)	79.4/3.5	81.4/4.5	75.2/6.9	**81.5/3.3**	78.9/5.2	80.2/3.9
Wine	94.8/2.6	94.6/3.4	88.1/4.6	87.4/6.8	93.9/3.9	93.9/4.1

Tabela 5. 27: Klasificijska točnost in pripadajoči SD za k-NN klasifikator. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	88.0/1.7	**88.6/1.6**	87.4/1.6	88.0/1.7	88.1/1.5	83.7/3.5	86.1/2.5
Annealing	98.1/0.8	97.4/1.0	**99.0/0.6**	98.2/1.1	98.5/0.7	96.9/0.9	97.0/0.7
Breast tissue	83.8/3.6	86.7/4.3	84.0/4.2	86.2/3.0	**87.4/2.7**	84.8/3.4	86.4/3.6
Credit app.	90.6/1.1	91.9/1.4	91.2/1.4	91.0/1.4	**92.0/1.2**	91.3/1.1	90.4/1.1
Glass	82.9/5.8	83.4/8.3	80.2/5.9	85.1/6.3	86.5/6.8	82.2/7.0	82.1/8.1
Heart cleve.	**69.9/2.8**	69.2/3.8	65.8/4.9	69.0/3.1	69.7/3.2	69.2/4.0	69.1/4.3
Ionosphere	92.8/2.6	92.9/2.0	92.4/2.1	93.7/2.0	93.1/2.1	91.3/6.0	93.1/2.3
Iris	98.8/1.1	98.3/1.0	98.8/1.1	99.1/0.9	98.8/1.0	98.1/2.4	**99.5/0.7**
Statlog(Heart)	**90.4/1.9**	90.2/2.1	88.7/1.9	89.5/2.0	**90.4/2.0**	87.5/3.3	89.3/2.5
Wine	99.6/0.3	**99.4/0.4**	97.9/0.7	**99.4/0.4**	**99.4/0.7**	99.1/0.7	98.7/1.4

Tabela 5. 28: AUC in pripadajoči SD za naivni Bayesov klasifikator. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	86.1/2.7	86.8/3.9	79.8/11.0	84.5/3.3	87.7/1.6	87.8/1.8
Annealing	97.7/0.9	97.8/1.0	73.8/25.05	92.7/15.0	97.4/1.1	97.3/1.1
Breast tissue	86.5/2.8	85.7/2.3	82.2/3.4	82.1/4.5	85.6/2.4	85.9/4.4
Credit app.	91.7/1.2	91.5/1.5	87.1/13.1	91.1/1.2	91.5/1.6	91.5/1.3
Glass	**87.9/6.8**	85.5/6.0	71.8/12.9	79.2/6.5	85.6/5.0	84.8/7.5
Heart cleve.	68.6/3.2	68.9/3.6	62.7/9.5	67.7/7.2	68.6/2.9	69.4/3.2
Ionosphere	93.1/2.4	**97.0/4.2**	80.5/10.8	85.5/4.9	92.9/1.0	89.0/2.8
Iris	99.2/0.8	98.8/1.1	85.4/10.2	78.0/10.1	99.2/1.0	99.4/0.9
Statlog(Heart)	89.5/2.5	90.1/1.8	79.8/11.1	85.6/3.4	89.3/2.2	89.6/1.8
Wine	**99.4/0.5**	99.2/0.6	95.3/4.9	94.3/5.6	99.1/0.9	99.1/0.6

Tabela 5. 29: AUC in pripadajoči SD za naivni Bayesov klasifikator. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	77.0/7.2	76.2/9.2	51.0/2.4	77.4/12.3	76.3/6.6	75.4/7.7	**79.2/2.7**
Annealing	95.8/3.4	94.4/2.7	95.4/3.4	96.4/2.6	96.5/2.7	96.7/1.7	94.3/3.5
Breast tissue	81.9/4.9	83.3/4.0	71.3/6.4	82.4/4.1	81.0/4.6	80.9/4.6	**84.3/3.3**
Credit app.	87.4/3.2	88.8/4.2	83.1/5.3	**90.0/3.4**	87.8/4.8	88.1/3.6	88.3/4.0
Glass	78.4/6.5	81.7/6.8	63.6/5.8	82.1/5.4	74.5/6.6	**82.4/6.2**	81.8/7.4
Heart cleve.	61.0/4.9	**61.6/6.2**	57.4/7.7	59.8/5.3	58.8/5.0	59.2/4.0	60.9/5.2
Ionosphere	86.8/4.6	89.9/5.5	80.2/4.1	87.7/4.1	89.2/4.4	88.2/4.0	89.0/4.3
Iris	98.0/1.5	95.7/2.4	97.8/0.7	97.7/0.8	**98.7/0.8**	97.8/1.9	98.6/0.6
Statlog(Heart)	79.2/6.9	80.0/5.5	79.9/4.4	80.8/3.2	80.2/5.7	77.3/6.7	79.4/3.9
Wine	95.4/2.1	96.3/1.8	85.6/5.7	95.6/2.3	95.5/2.4	96.1/2.4	93.6/4.4

Tabela 5. 30: AUC in pripadajoči SD za odločitvena drevesa. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	69.7/16.7	72.0/14.3	76.4/10.0	74.4/8.9	76.5/6.3	78.5/3.8
Annealing	97.1/1.7	97.3/1.7	72.5/23.8	89.6/14.3	**97.6/1.9**	95.7/3.4
Breast tissue	81.6/3.7	83.3/2.8	81.4/5.7	83.5/3.8	81.4/7.4	82.5/6.3
Credit app.	89.3/2.3	89.6/1.5	85.0/12.5	89.1/2.9	88.1/4.8	87.0/4.1
Glass	76.9/9.9	75.9/7.3	72.0/12.9	79.2/7.8	80.2/7.7	81.6/7.2
Heart cleve.	61.2/4.9	60.1/5.3	55.7/6.0	58.3/4.4	60.3/6.9	60.9/5.1
Ionosphere	**90.4/3.1**	89.2/4.4	79.4/11.0	85.3/5.0	87.8/2.1	87.2/3.9
Iris	98.2/1.1	98.2/1.1	86.6/9.4	79.0/9.4	98.3/1.1	98.5/0.6
Statlog(Heart)	79.6/5.8	79.8/4.2	77.0/11.0	**81.9/4.0**	77.8/5.7	78.6/4.6
Wine	96.0/2.0	**96.6/2.4**	93.1/3.9	93.3/5.4	93.7/4.7	93.8/4.6

Tabela 5. 31: AUC in pripadajoči SD za odločitvena drevesa. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	87.1/1.3	87.6/1.1	75.1/2.1	**88.7/1.2**	88.2/1.3	84.6/2.1	86.2/2.0
Annealing	99.4/0.2	98.6/0.9	99.3/0.5	99.1/1.0	**99.5/0.3**	99.0/0.4	98.9/0.6
Breast tissue	87.5/3.7	90.0/3.2	85.0/4.7	88.2/3.4	89.8/2.0	87.9/2.9	**90.5/3.3**
Credit app.	91.0/1.0	92.5/1.2	91.1/1.8	92.0/1.5	**92.7/1.1**	92.1/1.1	91.4/1.1
Glass	88.6/5.5	91.7/3.3	83.6/5.0	89.6/3.2	91.7/3.5	88.5/6.5	89.6/2.7
Heart cleve.	**67.6/3.3**	65.9/4.3	64.9/4.5	65.1/3.9	66.0/3.7	65.8/4.6	65.6/5.7
Ionosphere	**97.4/1.3**	96.0/1.9	92.7/2.3	96.1/1.1	96.4/1.3	95.2/2.6	95.9/1.7
Iris	99.5/0.5	99.5/0.4	99.7/0.2	99.2/0.8	99.3/0.4	99.6/0.4	**99.8/0.2**
Statlog(Heart)	**89.8/2.7**	**89.8/2.4**	86.3/2.6	89.3/2.0	89.2/2.5	86.5/3.6	89.0/2.3
Wine	99.0/0.2	99.7/0.2	99.2/0.7	99.7/0.4	99.7/0.4	99.3/0.6	99.2/0.6

Tabela 5. 32: AUC in pripadajoči SD za naključne gozdove. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	86.4/1.2	87.3/1.7	81.0/11.1	85.0/1.9	87.9/1.3	87.7/1.5
Annealing	99.3/0.3	99.4/0.2	74.4/25.71	94.1/15.5	99.1/0.3	99.2/0.6
Breast tissue	89.2/2.7	89.5/1.4	85.0/3.8	85.3/3.0	88.9/2.0	89.7/3.8
Credit app.	92.4/0.6	92.2/1.2	86.8/13.0	91.2/1.3	92.6/1.1	92.2/0.9
Glass	**92.8/4.0**	91.9/3.4	74.4/14.7	82.8/8.6	89.9/4.8	90.6/3.5
Heart cleve.	67.1/4.0	66.3/3.9	58.7/8.0	62.1/6.6	65.0/4.4	66.6/2.6
Ionosphere	96.3/1.8	96.1/2.4	81.5/11.2	89.4/4.8	96.4/1.6	94.5/2.1
Iris	**99.8/0.3**	99.7/0.3	87.0/9.4	79.0/9.6	**99.8/0.4**	99.7/0.3
Statlog(Heart)	88.5/ 2.9	89.1/2.1	81.6/11.8	87.2/1.7	87.4/3.3	89.2/2.2
Wine	**99.8/0.2**	99.7/0.3	96.7/3.9	95.4/3.2	99.3/0.7	99.3/0.7

Tabela 5. 33: AUC in pripadajoči SD za naključne gozdove. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	85.2/2.3	86.9/2.6	86.1/2.1	**86.7/2.2**	84.9/3.2	81.6/4.2	83.9/3.3
Annealing	97.0/2.5	96.6/2.8	**97.2/2.7**	95.5/2.8	96.9/2.5	96.7/2.1	95.9/2.6
Breast tissue	83.5/3.4	86.7/2.2	84.1/3.1	85.7/4.6	**88.0/2.4**	85.6/2.6	87.1/3.2
Credit app.	89.1/1.5	**91.7/1.1**	90.9/1.2	90.8/1.8	91.3/1.2	90.1/1.5	89.6/1.5
Glass	80.4/5.5	85.9/5.7	78.7/6.8	83.1/5.4	87.9/4.1	82.9/8.8	81.7/7.7
Heart cleve.	63.4/5.4	64.7/3.5	61.9/4.2	62.9/4.2	62.5/5.6	63.4/6.4	64.5/5.6
Ionosphere	94.2/2.2	95.2/2.8	91.5/1.6	91.4/3.1	**95.8/1.4**	93.8/3.6	92.4/3.0
Iris	98.7/1.2	99.1/0.8	99.1/0.5	99.0/0.8	98.7/0.8	99.2/1.0	99.5/0.8
Statlog(Heart)	**89.2/3.5**	88.5/2.9	87.7/2.2	88.3/1.9	88.4/2.7	85.8/2.9	87.7/2.9
Wine	99.5/0.5	99.7/0.4	**99.8/0.2**	99.4/0.7	99.6/0.5	99.1/0.9	98.4/2.2

Tabela 5. 34: AUC in pripadajoči SD za k-NN klasifikator. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	84.3/3.3	84.9/4.4	78.4/10.9	83.1/3.4	85.3/3.7	85.2/3.3
Annealing	96.7/2.7	96.8/2.0	71.9/23.1	92.1/15.0	97.0/2.6	96.7/2.4
Breast tissue	86.6/3.1	86.1/2.5	82.3/5.2	84.2/3.5	87.5/2.6	86.5/3.6
Credit app.	90.8/1.0	91.0/1.4	85.6/12.6	89.3/1.4	90.9/1.3	90.7/1.0
Glass	**89.1/4.7**	87.7/5.3	67.5/11.0	77.8/8.2	86.0/5.7	84.6/5.3
Heart cleve.	65.0/5.1	64.5/5.9	58.2/7.1	62.5/5.2	62.7/5.2	**67.8/2.9**
Ionosphere	93.1/1.1	92.5/4.4	80.4/11.3	88.5/4.6	94.1/4.0	90.1/4.1
Iris	99.6/0.7	99.4/0.8	86.0/10.1	78.4/9.4	99.5/0.6	**99.8/0.2**
Statlog(Heart)	88.6/2.3	88.2/2.9	82.6/11.8	88.5/3.4	88.3/2.5	88.7/2.4
Wine	99.5/0.5	99.7/0.3	96.6/2.9	94.4/5.5	99.4/0.6	99.5/0.5

Tabela 5. 35: AUC in pripadajoči SD za k-NN klasifikator. (2. del)

Domena	Ekvidist.	En. frek.	Požrešna	Fay.-Ir.	ReliefF	K	H
Adult	36	27.9	237.6	16	51.8	3.9/15.8	4.5/18.3
Annealing	36	25.9	105	25.3	57	4.3/18.1	4.0/19.2
Breast tissue	54	54	420	36.5	100.6	7.8/64.5	8.4/56.9
Credit app.	36	34	593.2	17.7	56.9	4.4/20.8	4.0/20.9
Glass	54	49.2	538.3	28	99.8	5.8/40.8	7.0/38.9
Heart cleve.	30	30	284.7	12	51.4	3.9/16.6	4.5/21.2
Ionosphere	201	198	1722	155.2	334.3	6.4/146.9	6.6/134.4
Iris	24	23.9	57.5	14.7	32.9	4.3/22.9	4.2/23.1
Statlog(Heart)	36	34.4	235.7	15.7	50.9	4.4/19.9	3.8/23.6
Wine	78	78	576.2	47	127.7	3.1/44.9	4.4/58.3

Tabela 5. 36: Prikaz števila dobljenih skupin končnih mej. (1. del)

Domena	Ks	Hs	Kcv	Hcv	EK	EH
Adult	4.3/21.5	3.3/21.5	3.0/13.1	5.8/14.9	22.4/28.5	26.1/30.1
Annealing	4.1/26.1	5.3/28.8	6.9/13	5.0/13.4	13.2/23.2	16.5/25.5
Breast tissue	3.8/65.3	3.4/68.2	7.2/28	7.6/24	9.9/70.9	10.3/60.1
Credit app.	3.5/26.1	5.0/31.7	7.7/14.9	6.3/15.5	12.1/32.3	10.1/32.8
Glass	3.4/55.5	4.2/53.9	5.2/22.6	7.6/24.6	8.0/47.8	7.9/41.9
Heart cleve.	2.8/31.8	5.2/31	3.1/11.3	3.9/12.3	11.9/30.7	11.3/30.2
Ionosphere	3.5/175.5	2.9/158.3	5.6/73.9	7.5/5.6	6.9/163.3	5.2/110.1
Iris	2.9/28.9	4.1/29.6	2.3/9.9	3.6/9.4	5.4/24.2	6.2/26.1
Statlog(Heart)	8.1/37.3	3.5/26.7	5.6/14	5.7/13.7	12.2/36.5	11.6/37.8
Wine	4.9/86.1	3.6/88.1	7.3/32.6	6.1/30.5	4.9/57.1	5.4/67.2

Tabela 5. 37: Prikaz števila dobljenih skupin končnih mej. (2. del)

Iz samih rezultatov, ki so v tabelah 5.20 do 5.35, je težko razbrati, kako se metode primerjajo med seboj. Zaradi tega smo vse rezultate rangirali in povprečja rangov prikazali v tabeli 5.38 in 5.39. Kasneje smo tudi s Friedmanovim testom preverili, če se metode med seboj statistično dovolj razlikujejo, da lahko rečemo, da so nekatere boljše oz. slabše od drugih.

Iz tabel 5.20 do 5.35 je razvidno, da se pri različicah Kcv in Hcv ponekod pojavlja bistveno večja standardna deviacija kot pri ostalih diskretizacijah. Zaradi tega metodi nista zanesljivi in ju lahko smatramo za neuporabni.

Tabeli 5.36 in 5.37 prikazujeta izbrano število skupin (samo pri diskretizacijah z razvrščanjem) in povprečno število uporabljenih mej. Opomnimo naj, da sta za vsak atribut izbrani tudi meji –Inf in Inf, tako da povprečno število intervalov dobimo tako, da vrednostim odštejemo 1.

Iz tabel 5.36 in 5.37 vidimo, da različici EK in EH v povprečju generirata več skupin, kot ostale različice diskretizacije z razvrščanjem. Tudi v primeru, kot je pri domeni Adult in različici diskretizacije EH, kjer je povprečno izbranih 26.1 skupin, metoda še vedno zadovoljivo deluje, čeprav hevristika nima lepe naraščajoče oblike, kot je prikazana na sliki 4.2, ker pri 30 rojih še ni prišlo do preloma. Razlog da metoda deluje tudi v takih primerih je v tem, da se pri večanju števila skupin, vedno bolj približujemo rezultatom, kot bi jih dala diskretizacija brez predhodnega razvrščanja, ki je uporabljena za izbiro končnih mej. Na primer, če bi izbrali število skupin enako številu učnih primerov, bi dobili vse možne meje (seveda, če pri tem ne uporabimo odstranjevanja majhnih intervalov, kot je opisano v razdelku 4.3).

	Ekv	**Fre.**	**Pož.**	**FI**	**Rel.**	**K**	**H**	**Ks**	**Hs**	**Kcv**	**Hcv**	**EK**	**EH**
Bayes	5.10	5.70	8.20	4.90	**4.25**	8.45	7.10	5.75	5.90	12.40	10.90	6.00	6.35
O. dre.	8.00	5.80	8.85	4.65	**5.15**	7.25	6.90	6.35	5.80	11.10	8.60	5.80	6.75
RF	5.10	**4.15**	8.20	5.30	4.40	8.50	7.30	6.30	6.05	12.50	10.95	6.85	5.40
k-NN	6.55	**4.75**	6.00	5.50	**4.75**	8.85	7.65	6.75	4.80	12.10	10.20	6.40	7.15
Skupaj	**6.19**	**5.10**	**7.81**	**5.09**	***4.64***	**8.26**	**5.46**	**6.29**	**5.64**	**12.03**	**10.16**	**6.26**	**6.41**

Tabela 5. 38: Povprečni rangi diskretizacij pri opazovanju klasificijske točnosti.

Izboljšana Friedmanova statistika F_F je za vse primere skupaj pri gledanju klasificijske točnosti 8.76. Kritična vrednost, da lahko zavrnemo ničelno hipotezo je 1.77. Iz tega sledi, da so si nekatere diskretizacije med seboj različne. Kritična razlika je 2.89 pri verjetnostjo napake prvega reda $p = 0.05$.

	Ekv.	**Fre.**	**Pož.**	**FI**	**Rel.**	**K**	**H**	**Ks**	**Hs**	**Kcv**	**Hcv**	**EK**	**EH**
Bayes	5.65	4.70	8.65	5.05	**2.65**	9.25	7.50	4.95	5.55	12.70	11.80	6.55	6.00
O. dre.	7.20	**5.05**	11.15	5.10	6.65	7.20	5.15	5.45	5.15	11.55	8.20	6.60	6.55
RF	6.35	5.10	11.05	6.40	**3.60**	8.70	7.10	4.05	4.50	12.65	11.50	5.60	5.40
k-NN	6.90	**3.95**	7.05	7.75	4.85	8.45	7.95	4.75	5.35	12.80	10.90	4.95	5.35
Skupaj	**6.51**	**4.69**	**9.46**	**6.07**	***4.43***	**8.39**	**6.92**	**4.83**	**5.18**	**12.43**	**10.60**	**5.95**	**5.81**

Tabela 5. 39: Povprečni rangi diskretizacij pri opazovanju AUC.

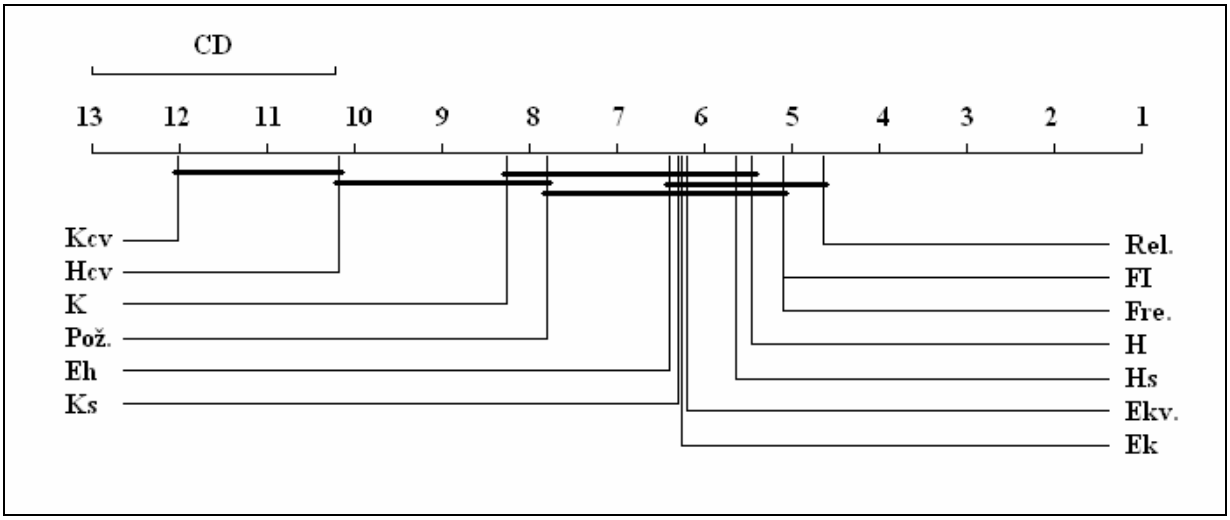
Tudi pri opazovanju skupnih rezultatov z AUC lahko s Friedmanovim testom zavrnemo ničelno hipotezo. Izboljšana statistika $F_F = 29.12$, kritična vrednost je tudi tukaj 1.77 ter enaka ostane tudi kritična razlika in sicer 2.89 pri verjetnosti napake prvega reda $p = 0.05$.

Iz tabel 5.38 in 5.39, da se različici Kcv in Hcv lahko primerjata le med seboj, od ostalih pa sta slabši, čeprav se je ravno ta metoda najboljše izkazala na umetnih primerih. Ti dve različici sta tudi najbolj časovno zahtevni, zaradi dvojne uporabe prečnega preverjanja.

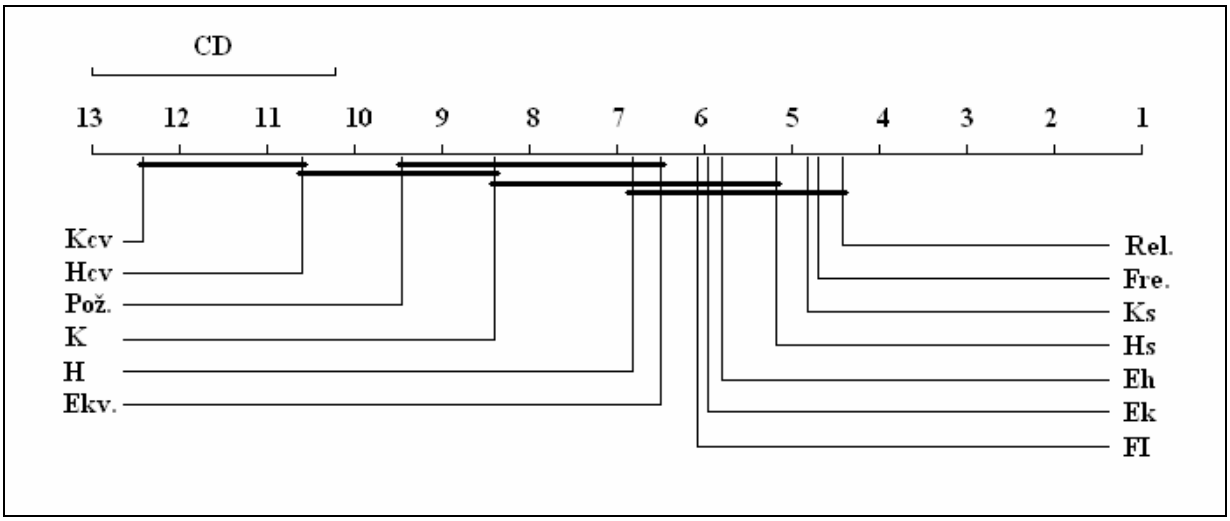
Pri opazovanju rangov klasificijske točnosti ima najboljši povprečni rang požrešna metoda z ReliefF-om in sicer 4.64, če z njo primerjamo druge diskretizacije lahko ugotovimo, da so požrešna z informacijskim prispevkom, ter različice diskretizacij z razvrščanjem K, Kcv ter Hcv statistično slabše, ker imajo rang nad 7.53(4.64+2.89). Podobno primerjavo naredimo pri opazovanju rangov AUC in pridemo do enakih zaključkov.

Vidimo, da so različice Ks, Hs, EK ter EH rangirane nekoliko bolje pri opazovanju AUC kot pri opazovanju klasificijske točnosti, a so razlike premajhne, da bi lahko karkoli z gotovostjo trdili.

Iz slik 5.4 in 5.5 naenkrat vidimo primerjave vseh diskretizacij. Skupine diskretizacij, ki niso statistično boljše/slabše so povezane z odebeljeno črto. Vse diskretizacije, ki so levo od skupine so slabše, vse desno pa boljše pri $p = 0.05$.



Slika 5. 4: Primerjava diskretizacij z Nemenyivim testom na podlagi klasificijske točnosti.



Slika 5. 5: Srimerjava diskretizacij z Nemeyivim testom na podlagi AUC.

Poglavje 6

6 Zaključek

V diplomskem delu smo razvili nove algoritme za diskretizacijo z uporabo razvrščanja. Za rešitev je bilo treba ugotoviti, kako povezati algoritme razvrščanja z obstoječimi diskretizacijami.

Pokazali smo, da lahko diskretizacija z razvrščanjem deluje bolje od obstoječih na umetnih domenah z močnimi odvisnostmi, ter da se vsaj nekatere od predlaganih različic zadovoljivo obnesejo tudi na realnih množicah.

Pri različici z prečnim preverjanjem za določitev števila skupin in prečnim preverjanjem za določitev končnih mej iz kandidatov se je pokazalo, da čeprav ta metoda deluje najbolje na umetnih domenah, to še ni zagotovilo, da se bo dobro obnesla tudi na realnih domenah.

Obstajajo tudi še možnosti za izboljšave in razširitve saj bi lahko preizkusili še druge vrste razvrščanja. Tako imenovano nadzorovano razvrščanje, bi najverjetneje lahko izboljšali z algoritmom, ki bi znal določiti različno število skupin za različne razrede učnih primerov. Kot smo videli v razdelku 5.2, bi bila potrebna izboljšava zaznavanja neuporabnih atributov, kar bi lahko dosegli z uteženo normalizacijo, pri čemer bi lahko neuporabnim atributom dali bistveno manjši razpon kakor tistim, za katere menimo, da so uporabni. Na primer pri problemu iz razdelka 5.2, bi atributa X in Y normalizirali na območje 0 do 100, atribut Z pa na 0 do 1. S tem postopkom bi razvrščanje dobilo podobne skupine kot v razdelku 5.1.

Literatura

[1] J. Demšar: Statistical Comparisons of Classifiers over Multiple Data Sets. V zborniku *Journal of Machine Learning Research 7*, 2006

[2] U.M. Fayyad, K.B.Irani: Multi-interval discretization of continuous-valued attributes for classification learning, V zborniku *The 13th International Joint Conference on Artificial Inteligence*, Morgan Kaufmann, pp. 1022-1027

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

[4] M. Robnik-Šikonja, I. Kononenko: Discretization of continuous attributes using ReliefF. V zborniku *ERK'95*, Portorož, Slovenija, 1995.

[5] Cluster analysis – Wikipedia, the free encyclopedia. Dostopno na: http://en.wikipedia.org/wiki/Cluster_analysis

[6] Determening the number of clusters in a data set – Wikipedia, the free encyclopedia. Dostopno na: http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[7] Hierarchial clustering – Wikipedia, the free encyclopedia. Dostopno na: http://en.wikipedia.org/wiki/Hierarchical_clustering

[8] k-means clustering – Wikipedia, the free encyclopedia. Dostopno na: http://en.wikipedia.org/wiki/K-means_clustering

[9] UCI Machine Learning Repositry: Data Sets. Dostopno na: http://archive.ics.uci.edu/ml/datasets.html