UNIVERZITET U NOVOM SADU
TEHNIČKI FAKULTET „MIHAJLO PUPIN“
ZRENJANIN

APROKSIMATIVNA DISKRETIZACIJA TABELARNO ORGANIZOVANIH PODATAKA

APPROXIMATIVE DISCRETIZATION OF TABLE-ORGANIZED DATA

Doktorska disertacija

kandidat
mr Višnja Ognjenović

Zrenjanin, 2016. godina

UNIVERZITET U NOVOM SADU
TEHNIČKI FAKULTET „MIHAJLO PUPIN“
ZRENJANIN

APROKSIMATIVNA DISKRETIZACIJA TABELARNO ORGANIZOVANIH PODATAKA

APPROXIMATIVE DISCRETIZATION OF TABLE-ORGANIZED DATA

Doktorska disertacija

mentor
Prof. dr Vladimir Brtka

kandidat
mr Višnja Ognjenović

Zrenjanin, 2016. godina

UNIVERZITET U NOVOM SADU
TEHNIČKI FAKULTET „MIHAJLO PUPIN“
ZRENJANIN

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: RBR	
Identifikacioni broj: IBR	
Tip dokumentacije: TD	Monografska dokumentacija
Tip zapisa: TZ	Tekstualni štampani materijal
Vrsta rada: VR	Doktorska disertacija
Autor: AU	mr Višnja Ognjenović
Mentor/Ko-mentor: MN	Prof. dr Vladimir Brtka
Naslov rada: NS	Aproksimativna diskretizacija tabelarno organizovanih podataka
Jezik publikacije: JZ	Srpski (Latinica)
Jezik izvoda: JI	Srpski/Engleski
Zemlja publikovanja: ZP	Republika Srbija
Uže geografsko područje: UGP	AP Vojvodina
Godina: GO	2016.
Izdavač: IZ	Autorski reprint
Mesto i adresa: MS	Zrenjanin, Đure Đakovića bb
Fizički opis rada: broj poglavlja/broj strana/broj slika (ilustracija)/broj tabela/broj referenci FO	8/165/232/18/101
Naučna oblast: OB	Informacione tehnologije
Naučna disciplina: ND	Mašinsko učenje, Data mining

Predmetna odrednica, ključne reči: PO	diskretizacija podataka, aproksimativna diskretizacija, klasifikacija, MD heuristics, tabelarno organizovni podaci, segmentacija multimodal raspodele, konzistentna\inkonzistentna tabela odlučivanja, teorija grubih skupova, raspodela podataka
UDK	
Čuva se: ČU	Biblioteka Tehničkog fakulteta „Mihajlo Pupin“ Zrenjanin
Važna napomena: VN	nema
Izvod: IZ	Disertacija se bavi analizom uticaja raspodela podataka na rezultate algoritama diskretizacije u okviru procesa mašinskog učenja. Na osnovu izabranih baza i algoritama diskretizacije teorije grubih skupova i stabala odlučivanja, istražen je uticaj odnosa raspodela podataka i tačaka reza određene diskretizacije. Praćena je promena konzistentnosti diskretizovane tabele u zavisnosti od položaja redukovane tačke reza na histogramu. Definisane su fiksne tačke reza u zavisnosti od segmentacije multimodal raspodele, na osnovu kojih je moguće raditi redukciju preostalih tačaka reza. Za određivanje fiksnih tačaka konstruisan je algoritam FixedPoints koji ih određuje u skladu sa grubom segmentacijom multimodal raspodele. Konstruisan je algoritam aproksimativne diskretizacije APPROX MD za redukciju tačaka reza, koji koristi tačke reza dobijene algoritmom maksimalne razberivosti i parametre vezane za procenat nepreciznih pravila, ukupni procenat klasifikacije i broj tačaka redukcije. Algoritam je kompariran u odnosu na algoritam maksimalne razberivosti i u odnosu na algoritam maksimalne razberivosti sa aproksimativnim rešenjima za α=0,95.
Datum prihvatanja teme od strane NN veća: DP	
Datum odbrane: DO	

Članovi komisije: ČK	predsednik: Prof. dr Ivana Berković, redovni profesor, Univerzitet u Novom Sadu, Tehnički fakultet „Mihajlo Pupin“, Zrenjanin član: Doc. dr Željko Stojanov, docent, Univerzitet u Novom Sadu, Tehnički fakultet „Mihajlo Pupin“, Zrenjanin član: Doc. dr Dalibor Dobrilović, docent, Univerzitet u Novom Sadu, Tehnički fakultet „Mihajlo Pupin“, Zrenjanin član: Prof. dr Milena Stanković, , redovni profesor, Univerzitet u Nišu, Elektronski fakultet, Niš član - mentor: Prof. dr Vladimir Brtka, vanredni profesor, Univerzitet u Novom Sadu, Tehnički fakultet „Mihajlo Pupin“, Zrenjanin

UNIVERSITY OF NOVI SAD
TECHNICAL FACULTY “MIHAJLO PUPIN”
ZRENJANIN

KEY WORD DOCUMENTATION

Accession number: ANO	
Identification number: INO	
Document type: DT	Monographic publication
Type of record: TR	Textual material, printed
Concents code: CC	Ph.D. Thesis
Author: AU	Višnja Ognjenović, M.Sc.
Menthor: MN	Vladimir Brtka, PhD, associate professor.
Title: TI	Approximative Discretization of Table-Organized Data
Language of text: LT	Serbian (Latin letters)
Language of abstract: LA	Serbian and English
Contry of publication: CP	Serbia
Locality od publication: LP	Vojvodina
Publication year: PY	2016.
Publisher: PU	Author reprint
Publication place: PP	Zrenjanin, Đure Đakovića bb
Physical description (chapters/pages/images (illustrations)/tables/pages): PD	8/165/232/18/101
Scientific field: SF	Information Technologies
Scientific discipline: SD	Machine learning, Data mining

Subject, Key words: SKW	data discretization, approximate discretization, classification, MD heuristics, tabular data - tabular presentation of data, multimodal segmentation, consistent\inconsistent decision tables, rough set theory, data distribution
UDC	
Holding data: HD	Library of Technical faculty “Mihajlo Pupin”, Zrenjanin
Note: N	none
Abstract: AB	This dissertation analyses the influence of data distribution on the results of discretization algorithms within the process of machine learning. Based on the chosen databases and the discretization algorithms within the rough set theory and decision trees, the influence of the data distribution-cuts relation within certain discretization has been researched. Changes in consistency of a discretized table, as dependent on the position of the reduced cut on the histogram, has been monitored. Fixed cuts have been defined, as dependent on the multimodal segmentation, on basis of which it is possible to do the reduction of the remaining cuts. To determine the fixed cuts, an algorithm FixedPoints has been constructed, determining these points in accordance with the rough segmentation of multimodal distribution. An algorithm for approximate discretization, APPROX MD, has been constructed for cuts reduction, using cuts obtained through the maximum discernibility (MD-Heuristic) algorithm and the parametres related to the percent of imprecise rules, the total classification percent and the number of reduction cuts. The algorithm has been compared to the MD algorithm and to the MD algorithm with approximate solutions for α=0,95.
Accepted on Scientific Board on: AS	
Defended: DE	

Thesis Defend Board: DB	President: Ivana Berković, PhD, full professor, University of Novi Sad, Technical faculty “Mihajlo Pupin”, Zrenjanin Member: Željko Stojanov, PhD, assistant professor , University of Novi Sad, Technical faculty “Mihajlo Pupin”, Zrenjanin Member: Dalibor Dobrilović, PhD, assistant professor , University of Novi Sad, Technical faculty “Mihajlo Pupin”, Zrenjanin Member: Milena Stanković, PhD, full professor, University of Niš, Faculty of Electronic Engineering, Niš Member, Mentor: Vladimir Brtka, PhD, associate professor, University of Novi Sad, Technical faculty “Mihajlo Pupin”, Zrenjanin

mojoj Milici

Zahvalnost

„I Videlo se svetli u tami, i tama Ga ne obuze."

Jovan 1,5

Zahvaljujem što mi je omogućeno da radim i istražujem, da sarađujem sa dragim profesorima i kolegama, da otkrivam nove i korisne stvari.

Zahvaljujem mentoru Prof. dr Vladimiru Brtki na podsticaju i kritikama. Zahvaljujem Prof. dr Ivani Berković na podršci i pomoći. Zahvaljujem članovima komisije Prof. dr Mileni Stanković, Doc dr. Daliboru Dobriloviću i Doc. dr Željku Stojanovu na podršci i saradnji.

Zahvaljujem kolegama Doc. dr Eleonori Brtki, Doc. dr Vesni Makitan, Doc. dr Jeleni Stojanov, dr Ljubici Kazi, mr Gordani Jotanović, Doc. dr Zoltanu Kazi i dr Dejanu Lacmanoviću na podršci i uticaju.

Hvala strini Mili, stricu Dražimiru i bratu Martinu na inspiraciji. Posebno hvala mojoj mami Mileni, sestri Milici, svekrvi Ljiljani, suprugu Jovanu, ćerki Jani i sinu Vukoti što su bili uz mene.

mr Višnja Ognjenović

Aproksimativna diskretizacija tabelarno organizovanih podataka

Sadržaj

1 UVOD

Tema ove disertacije je iz oblasti Data Mining-a, u okviru rada sa tabelarno organizovanim podacima. U oblasti Data Mining-a mnogi obučavajući metodi (*machine learning*) mogu da rade samo sa diskretnim vrednostima atributa. Zbog toga je pre machine learning procesa, neophodna transformacija kontinualnih vrednosti atributa u diskretne, konstituisane od skupa intervala. Ovaj proces poznat kao diskretizacija podataka je esencijalni zadatak u preprocesiranju podataka, ne samo zbog zbog toga što pojedini obučavajući metodi ne rade sa kontunualnim vrednostima atributa, već i zbog toga što su podaci transformisani u skup intervala kognitivno relevantni za ljudska tumačenja. Rezultat diskretizacije podataka je skup tačaka kojima se podaci svrstavaju u intervale. U zavisnosti od konkretnog algoritma diskretizacije, koji je konceptualno vezan za određenu teoriju ili metod, razvijaju se metodi optimizacije algoritma, heuristike, a takođe i aproksimativne vrednosti. Aproksimativne vrednosti neke diskretizacije bazirane su na manjem skupu tačaka koje proizvode intervale, u odnosu na rezultat početne diskretizacije.

Postoji praktična potreba razvoja algoritama aproksimativne diskretizacije. Smanjenje broja tačaka može da omogući brži rad algoritma, kao i bolje rezultate klasifikacije. Naknadnim smanjivanjem broja tačaka, određeni automatizam nekog algoritma se može povezati i sa uticajem eksperta. Time se omogućuje kompleksniji rad algoritma diskretizacije.

Empirijski rezultati pokazuju da kvalitet klasifikacijskih metoda zavisi od algoritma diskretizacije koji se koristi. Imajući u vidu da je diskretizacija proces traženja particija domena atributa i ujednačavanja vrednosti u okviru svih intervala, problem diskretizacije se može definisati kao problem traženja relevantnih skupova tačaka reza (*cuts*) nad domenima atributa. Kvalitet neke diskretizacije se često meri u odnosu na rezultat klasifikacije. Ako se za klasifikaciju koristi određena matematička teorija, onda je često „kompatibilno“ da se i podaci diskretizuju prema istim formulama.

Mnogi istraživači su primetili veliki uticaj koji diskretizacija podataka ima na rezultat klasifikacije. Za određeni algoritam klasifikacije rađene su komparacije mnogih algoritama za diskretizaciju. Na osnovu statističkih pokazatelja izvođeni su zaključci i razvijani novi algoritmi. Bez obzira na složenost nekog algoritma za diskretizaciju u odnosu na to da li se analiziraju svi podaci kao celina ili pojedinačni delovi, da li postoji povratna sprega u algoritmu ili neki zadati parametri, razvijani su metodi za dodatna poboljšanja poput redukcije određenih atributa ili određenih tačaka reza. Za neke algoritme diskretizacije definisani su i napravljeni algoritmi koji daju aproksimativna rešenja. Na osnovu preovlađujuće normalne raspodele podataka, napravljen je Approximate Equal Frequency algoritam za diskretizaciju. U okviru Teorije grubih skupova omogućen je rad odgovarajućeg algoritma za diskretizaciju sa aproksimativnim početnim izračunavanjima. Kombinovani su postojeći algoritmi u cilju optimizacije konačnog rezultata klasifikacije.

U okviru dosadašnjeg rada sa klasifikatorima, pre svega u Teoriji grubih skupova, primećeno je da sa jedne strane posmatarno konačni rezultat klasifikacije zavisi od same raspodele podataka. Neki istraživači su to u svojim radovima uglavnom navodili kao konstataciju, bez upuštanja u analizu raspodele podataka. Urađeno je jedno istraživanje na veštački generisanim podacima sa različitim raspodelama i nad njima je

izvršena komparacija određenog broja diskretizacija. To istraživanje nije nastavljeno na proizvoljnim podacima. Na osnovu razvijene matematike teorije grubih skupova, dokazano je da kod velikih baza određena tačka reza može da se izbaci u zavisnosti od medijane funkcije raspodele određenog atributa. Iskustvo rada sa raznim diskretizacijama navelo je na proveru zakonitosti između položaja tačaka reza na raspodeli podataka u odnosu na primenjeni algoritam diskretizacije. U određenoj meri to je urađeno u jednom delu istraživanja ove disertacije. Dobijeni rezultati su ugrađeni u metod redukcije tačaka reza na osnovu histograma.

Drugi ugao posmatranja diskretizacija je vezan za konzistentnost tabele sa podacima. Drugačije rečeno postojanje istih uslova koji daju različite odluke dovodi do procentualno bolje klasifikacije, ali u osnovi takva klasifikacija je vezana za neprecizne uslove klasifikovanja. Ovaj problem je analiziran u jednom delu istraživanja ove disertacije i ponuđen je algoritam koji omogućuje praćenje neprecizne klasifikacije.

Analiza raspodele podataka putem histograma i inkonzistentnost tabele podataka, omogućila je da se osmisli algoritam aproksimativne diskretizacije baziran na algoritmu diskretizacije maksimalne razberivosti, koji je nazvan APPROX MD. Time je analiziranje rezultata klasifikacije pored procenta klasifikovanih podataka, prošireno na praćenje unapred zadate gornje granice inkonzistentnosti dela diskretizovane tabele na osnovu procenta nepreciznih pravila.

Za istraživanje i komparaciju korišteno je deset javno dostupnih baza. One su birane tako da budu reprezentativne i analizirane su na osnovu svih ciljeva i zadataka ove disertacije.

Posmatrano po poglavljima, ovaj rad je podeljen u sledeće celine:

I. Prvo poglavlje je Uvod.

II. U drugom poglavlju je dat prikaz metodologije istraživanja sa naglaskom na hipoteze i pothipoteze.

III. U trećem poglavlju je dat prikaz osnovnih algoritama diskretizacije sa posebnom analizom diskretizacije bazirane na entropiji i diskretizacije povezane sa teorijom grubih skupova. Konkretno u trećem poglavlju je analizirana diskretizacija podataka za algoritam C4.5 i algoritam maksimalne razberivosti koji se koristi u Teoriji grubih skupova.

IV. Četvrto poglavlje daje pregled stanja iz područja istraživanja.

V. Istraživanje se nalazi u petom poglavlju i rađeno je na osnovu uticaja raspodele podataka na algoritme diskretizacije, analizu klasifikacija na osnovu raspodela podataka i razradu ideje redukcije tačaka reza dobijenih algoritmom maksimalne razberivosti.

VI. Glavni doprinos rada je prikazan u šestom poglavlju i predstavljen je metodom redukcije tačaka reza i samim algoritmom aproksimativne diskretizacije APPROX MD koji je baziran na algoritmu maksimalne razberivosti.

VII. U sedmom poglavlju je komparacija dobijenog aproksimativnog algoritma za diskretizaciju APPROX MD sa algoritmom maksimalne razberivosti i algoritmom maksimalne razberivosti sa aproksimativnim rešenjima za α=0,95.

VIII. U osmom poglavlju se nalazi zaključak i smernice budućih istraživanja.

Nakon osmog poglavlja je spisak korišćenih literaturnih izvora.

Na kraju rada se nalaze dodaci za analizirane baze: dobijene tačke reza i histogrami atributa analiziranih baza sa ucrtanim tačkama reza za svaki analizirani algoritam.

2 METODOLOGIJA ISTRAŽIVANJA

2.1 PREDMET ISTRAŽIVNJA

Predmet istraživanja je vezan na praktičnu i čestu radnju u informatici, konkretno u oblasti Data Mining [Han i dr., 2011], Teoriji grubih skupova [Pawlak, 1982] [Nguyen, 2006] [Grzymala-Busse, 2005] [Suraj, 2004], a to je klasifikacija i analiziranje rezultata klasifikacije [Bazan i dr., 2000]. U odnosu na klasifikaciju bitan je prvi korak, a za numeričke vrednosti podataka to je diskretizacija [Komorowski i dr., 1999] [Nguyen, 2006]. Nakon diskretizacije podaci se dele na trening i test skup; na osnovu trening skupa generišu se pravila odlučivanja i na osnovu pravila odlučivanja se podaci iz test skupa klasifikuju [Komorowski i dr., 1999] [Brtka i dr., 2011b]. Izazov je bio da se prvo prouče podaci, odnosno njihova raspodela podataka data histogramom i da se zatim posmatraju rezultati diskretizacije i ostali koraci do konačnih rezultata klasifikacije. Mali je broj radova koji su detaljno analizirali histograme podataka pri procesu diskretizacije a analizu su radili samo sa po jednog aspekta [Ismail i dr., 2003] [Kontkanen i dr., 2007] [Schmidberger i dr, 2005]. Stoga je ideja posmatranja histograma i tačaka reza (dobijenih diskretizacijom) postavljena kao primarni zadatak.

Ako se posmatraju tabelarni podaci (informaciona tabela, informacioni sistem) sa dve različite vrste atributa, uslovnih i atributa odluke, onda se za takvu tabelu koristi naziv tabela odlučivanja (*decision tables*). Svaka vrsta (red) tabele odlučivanja određuje pravila odlučivanja u formi IF THEN, kojom se utvrđuju odluke koje se preduzimaju pod uslovom da su zadovoljeni uslovni atributi [Pawlak, 2002] [Brtka, 2008].

S obzirom da svaka diskretizacija podataka u nekoj meri predstavlja gubitak (originalnih) podataka, analizirana je promena konzistentnosti tabelarnih podataka. Diskretizovani podaci dele tabelu na disjunktne podskupove. Ako dva različita elementa takvog podskupa imaju istu odluku, sistem je konzistentan; inače nije. U odnosu na relaciju ili operaciju kojom se radi diskretizacija podataka, ispitivanje konzistentnosti je aktuelan zadatak [Qian i dr., 2010 a] [Tang i dr., 2014] [Du i dr., 2015]. Da bi se konzistentnost posmatrala na što jednostavniji način, bez formula i različitih indeksa koji su definisani zadnjih godina, posmatrana je direktna implikacija inkonzistentne tabele: pravila odlučivanja koja u IF delu imaju iste uslove, a u THEN delu različite odluke. Važi i suprotno, naime dva pravila iz tabele odlučivanja su inkonzistenta akko imaju iste uslove a različite odluke [Lover, 2008]. Inkonzistenta (nederministička, konfliktna, moguća) pravila nam govore da se različite stvari događaju na osnovu istih uslova. Tabela odlučivanja koja sadrži inkonzistentna pravila odlučivanja se zove inkonzistentna; inače je konzistentna [Pawlak, 2001].

Broj konzistentnih pravila na prema broju svih pravila tabele odlučivanja može da se koristi kao faktor konzistentnosti tabele i označava se sa $\gamma(C,D)$, gde su C i D uslovni i atributi odluke respektivno [Polkowski, 2003]. Zbog manjeg broja inkonzistentnih pravila u odnosu na konzistentna, u ovoj disertaciji je utvrđivanje inkonzistentnosti tabele izabrano na osnovu procenta inkonzistentnih pravila. To je urađeno da bi rezultati istraživanja na jednostavniji način bili primenljiv u praksi. Zbog samog izgleda inkonzistentnih pravila, čijim sažimanjem mogu da se dobiju pravila koja u THEN delu imaju operator OR (na osnovu više odluka), u ovoj disertaciji će se koristiti i termin neprecizna pravila.

U cilju redukcije određenih tačaka reza, ideja je bila da se prati njihov položaj na histogramu, pripadnost reduktu skupa podataka, uticaj na klasifikaciju i na konzistentnost. Redukcija tačaka reza je postavljena kao neizostavni deo predmeta istraživanja, jer rešavanje ovog problema direktno vodi do mogućnosti pravljenja algoritma za aproksimativnu diskretizaciju.

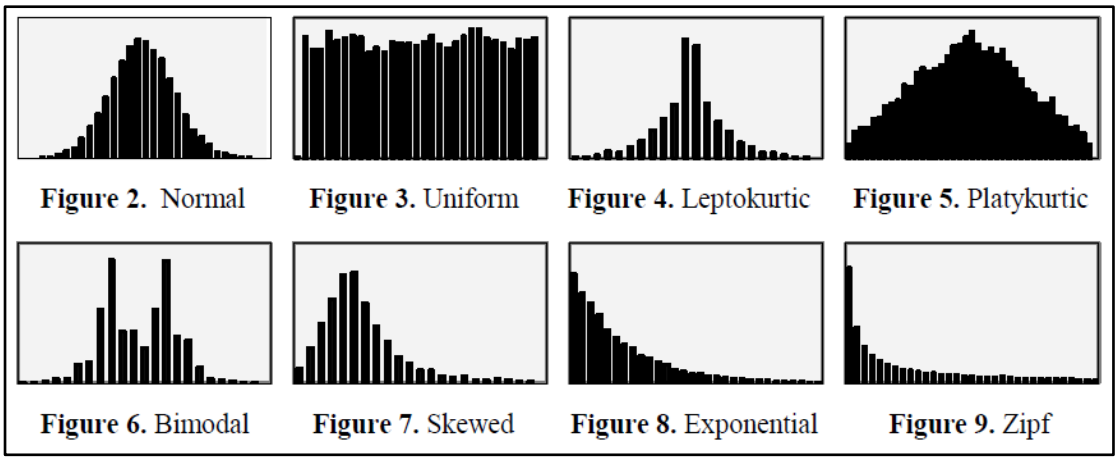
Posmatrano u celini predmet istraživanja je:

- odnos raspodele podataka i rezultata algoritama za diskretizaciju u okviru tabelarno organizovanih podataka;
- odnos rezultata klasifikacije i konzistentnosti tabelarno organizovanih podataka;
- mogućnost redukcije tačaka reza dobijenih algoritmom diskretizacije maksimalne razberivosti;
- sumiranje dobijenih rezultata u cilju pravljenja algoritma aproksimativne diskretizacije.

Diskretizacijom podataka dobijaju se tačke reza, koje je moguće analizirati na histogramu raspodele podataka [Ognjenović i dr., 2013]. Da bi se uradio eksperiment u određenom vremenu, potrebno je izabrati određen broj baza i određen broj algoritama za diskretizaciju. Na osnovu izvršene procene vremena i broja baza, i različitih algoritama koje su birali drugi istraživači, izabrano je deset baza i dva algoritma za diskretizaciju. Baze su izabrane tako da budu javno dostupne, da imaju jedan atribut odluke, da imaju različit broj atributa, da su različite po broju objekata, da imaju veći broj ili isključivo atribute sa numeričkim vrednostima i da im atributi imaju razne raspodele. Ovaj poslednji uslov je vezan za određenu meru sličnosti sa nekom matematičkom funkcijom.

Za jednostavan algoritam baziran na jednačenju broja objekata napravljen je aproksimativan algoritam Approximate Equal Frequency na osnovu (najčešće) normalne raspodele [Jiang i dr., 2009].

U radu koji je imao istraživanje na veštačko generisanim podacima [Ismail i dr., 2003], izabrano je osam raspodela koje su praćene - prikazane su na preuzetoj slici 1. Rezultati tog rada imaju zaključke za samo određene tipove raspodela u odnosu na izabrane algoritme diskretizacije, a algoritmi su komparirani samo na osnovu ukupnog rezultata klasifikacije.



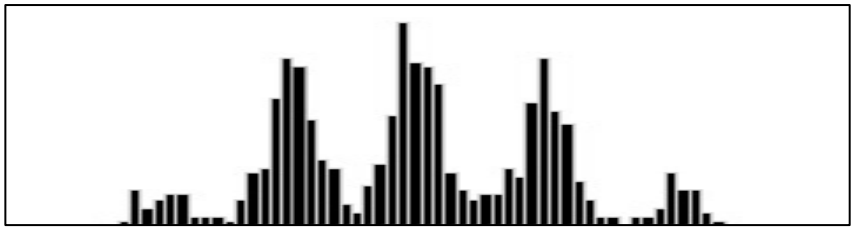
Slika 1. Raspodele koje su praćene u radu [Ismail, 2003]

U bazi koja nije veštački generisana izračunavanje sličnosti sa nekom konkretnom raspodelom je neproduktivno jer postoji čitav niz raspodela koje u određenoj meri odgovaraju histogramu stvarnih podataka. Dodatno, veza sa određenom funkcijom raspodele kod raspodela koje su kombinacija raspodela sa slike 1 bi bila veoma neprecizna.

Ono što je godinama ranije bilo primećeno vezano za položaj tačke reza za određeni algoritam diskretizacije u odnosu na raspodelu podataka jeste sledeće:

- kod algoritma za diskretizaciju koji je baziran na entropiji, u slučaju raspodele slične normalnoj (što je u stvari mesokurtic za k=0), uniformnoj, leptokurtic (k>0), platykurtic (k<0), skewed, exponential, zipf i ostalim raspodelama koje nemaju više izrazitih maksimuma (poput bimodal), dobija se jedan izrazito velik podinterval i nekoliko mnogo manjih;
- kod algoritma diskretizacije maksimalne razberivosti tačke reza prate okoline lokalnih maksimuma – u slučaju bimodal raspodele tačka reza bi verovatno bila u minimumu između dva maksimuma.

Zbog toga su sve raspodele podataka odnosno njihovi histogrami posmatrani sa dva stanovišta. Neke izabrane baze su na osnovu raspodele podataka birane tako da imaju sve ili značajan broj uslovnih atributa čija je raspodela slična sa normalnom, uniformnom, leptokurtic, platykurtic, skewed, exponential, zipf i ostalim raspodelama koje nemaju više izrazitih maksimuma. Druge baze su birane tako da imaju sve ili značajan broj uslovnih atributa čije raspodele su slične sa bimodal ili multimodal raspodelom sa proizvoljnim brojem lokalnih maksimuma kao na slici 2. Preostale baze su birane tako da približno polovina broja atributa pripada raspodelama sličnim sa normalnom, uniformnom, leptokurtic, platykurtic, skewed, exponential, zipf i ostalim raspodelama koje nemaju više izrazitih maksimuma, a druga polovina uslovnih atributa ima raspodelu sličnu sa bimodal ili multimodal raspodelom sa proizvoljnim brojem lokalnih maksimuma.



Slika 2. Primer multimodal raspodele

Od algoritama za diskretizaciju posmatraće se algoritam maksimalne razberivosti MD (*Maximal-Discernibility heuristics, MD-heuristics* ili kraće MD) [Nguyen, 1997] [Nguyen 2006] i algoritam baziran na entropiji [Dougherty i dr., 1995]. Za klasifikaciju diskretizovanih podataka koristiće se rezultati Džonsonovog algoritma [Johnson, 1974] za izračunavanje minimalnih prostih implikanti Bulove funkcije, odnosno na osnovu samo jednog redukta skupa generisaće se IF ... THEN pravila za klasifikovanje. Džonsonov algoritam je izabran zbog toga što daje jedan redukt. U slučaju generisanja većeg broja redukta problem bi se raširio tako da bi bilo teško uraditi mnogostruku komparaciju redukta, a to ne bi doprinelo rezultatu klasifikacije.

Na osnovu eksperimenta nad deset baza utvrdiće se na koji način raspodela podataka utiče na položaj tačke reza kod određenog izabranog algoritma za diskretizaciju. Dalje ispitivanje odnosa redukcije tačaka reza kod algoritma maksimalne razberivosti uradiće se na određenim bazama koje imaju reprezentativne atribute.

Dobijeni rezultati će se primeniti u kreiranju algoritma aproksimativne diskretizacije APPROX MD koji će predstavljati modifikaciju algoritma maksimalne razberivosti.

2.2 CILJEVI ISTRAŽIVANJA

Osnovni cilj ovog istraživanja je da se omogući modifikacija postojećih algoritama za diskretizaciju podataka tako da daju aproksimativne vrednosti na osnovu raspodela podataka. Konkretno, cilj je kreirati algoritam aproksimativne diskretizacije baziran na algoritmu diskretizacije maksimalne razberivosti tako da na osnovu raspodele podataka može da daje aproksimativna rešenja. U okviru ovog osnovnog cilja, mogu se definisati sledeći ciljevi:

Cilj I	Provera odnosa tačaka reza određenog algoritma diskretizacije i raspodele podataka. U okviru ove provere cilj je utvrđivanje učestalosti tačaka reza u odnosu na raspodelu podataka, kao i mesto tačaka reza u odnosu na lokalne ekstreme (prvenstveno maksimume) raspodele podataka date histogramom.
Cilj II	Provera odnosa raspodele podataka u odnosu na rezultat klasifikacije. Za ovaj cilj posmatraće se raspodele cele baze, dok će se rezultat klasifikacije posmatrati na osnovu matrice konfuzije.
Cilj III	Provera odnosa rezultata klasifikacije u odnosu na inkozistentnost tabele i algoritam diskretizacije.
Cilj IV	Uticaj redukcije broja tačaka reza na inkonzistentnost tabele. Cilj je ustanoviti koliko je rezultat klasifikacije bolji na osnovu matrice konfuzije, a koliko je bolji zbog inkonzistentnosti tabele.

2.3 ZADACI ISTRAŽIVANJA

U okviru definisanih ciljeva uradiće se sledeće:

Zadatak I — Za izabrane baze treba uraditi ispitivanje položaja tačke reza određenog algoritma diskretizacije u odnosu na raspodelu podataka. Za ovaj zadatak potrebno je posmatrati histogram svakog atributa svih izabranih baza. Za analizu histograma koristiće se softver EasyFit [Easy, 2015]. Crtanje tačaka reza na histogramu uradiće se u programu za vektorsku grafiku AdobeIllustrator. Nakon crtanja svih tačaka reza na svim histogramima jedne baze daće se odgovori na pitanje gde se očekuje tačka reza u odnosu na raspodelu podataka za algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji. Oba algoritma za diskretizaciju podataka će se koristiti u sistemu Rosetta [Øhrn, 1998].

Zadatak II — Proveriće se odnos raspodele podataka u odnosu na rezultat klasifikacije. Posmatraće se raspodela svih uslovnih atributa cele baze i rezultat klasifikacije dobijen u sistemu Rosetta na sledeći način:

- II.a. za trening skup će se uzeti polovina baze nad kojim će se generisati redukt na osnovu Džonsonovog algoritma;
- II.b. na osnovu generisanog redukta, generisaće se IF THEN pravila;
- II.c. na osnovu generisanih IF THEN pravila koja opisuju trening skup, uradiće se klasifikacija podataka druge polovine baze tako da će se kao rezultat klasifikacije dobiti matrica konfuzije;

Zadatak III — Provera odnosa rezultata klasifikacije u odnosu na inkozistentnost tabele i algoritam diskretizacije uradiće se tako što će se posmatrati ukupan procenat klasifikacije iz matrice konfuzije i procenat IF ... THEN pravila koja u THEN delu imaju operator OR (neprecizna pravila).

Zadatak IV — Pri redukciji tačaka reza, koliko je rezultat klasifikacije bolji na osnovu ukupnog procenta matrice konfuzije a koliko je bolji zbog inkonzistentnosti tabele? Rešavanje ovog zadatka treba da pomogne u kreiranju algoritma aproksimativne diskretizacije baziranog na algoritmu maksimalne razberivosti.

2.4 HIPOTEZE ISTRAŽIVANJA

Glavna hipoteza:

> Moguće je izgraditi algoritam za aproksimativne diskretizacije u odnosu na raspodele iz tabelarno organizovanih podataka.

Pothipoteze:

I Raspodele podataka utiču na algoritam diskretizacije.

II Raspodele podataka utiču na kvalitet klisifikacije.

III Diskretizacija podataka može značajno da utiče na inkonzistentnost tabele kod podataka sa određenim raspodelama.

IV Na osnovu raspodele podataka moguće je smanjiti broj tačaka diskretizacije tako da redukt skupa tabelarno organizovanih podataka ostane nepromenjen.

Bitan rezultat je modifikacija bar jednog algoritma za diskretizaciju podataka, tako da omogući rad sa aproksimativnim vrednostima na osnovu raspodela podataka.

3 ALGORITMI DISKRETIZACIJE

3.1 OSNOVNI ALGORITMI DISKRETIZACIJE

3.1.1 Naive algoritam

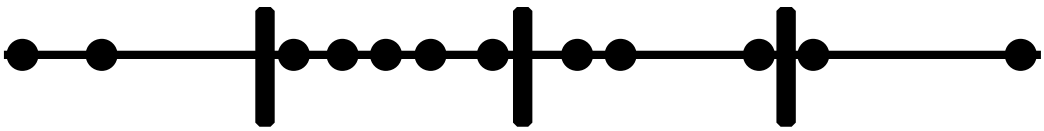
Naive algoritam diskretizacije je najjednostavniji od svih. Ako se posmatraju vrednosti jednog atributa, onda za sve njegove različite vrednosti, algoritam Naive generiše tačke reza kao aritmetičku sredinu susednih vrednosti. Slika 3 ilustruje rezultat rada ovog algoritma (vrednosti atributa su označene kružićima, a tačke reza vertikalnim linijama).

Slika 3. Naive algoritam

3.1.2 EqualWidth algoritam

EqualWidth (EW) ili algoritam jednakih podintervala je takođe jednostavan algoritam koji se izvršava nad vrednostima jednog atributa bez uticaja ostalih atributa. On generiše tačke reza tako da se početni interval deli na fiksan broj podintervala. U ovom slučaju korisnik mora da izabere željeni broj intervala.

U slučaju velikog broja podintervala može da dođe do problema obučavanja. Za premali broj podintervala postoji mogućnost gubljenja korisnih informacija [Muhlenbach i dr., 2005]. Na slici 4 čitav interval vrednosti atributa podeljen je na 4 podintervala jednake „širine“ (vrednosti atributa su označene kružićima, a tačke reza vertikalnim linijama).

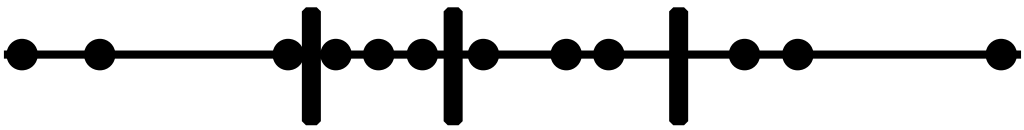


Slika 4. EqualWidth algoritam

3.1.3 Equal frequency algoritam

Equal frequency algoritam (EF) je takođe jednostavan algoritam koji se izvršava nad vrednostima jednog atributa bez uticaja ostalih. Za unapred dat fiksan broj n željenih intervala, na osnovu histograma se određuje n-1 tačka reza tako da u svakom od n podintervala bude približno jednak broj objekata – idealno bi bilo 1/n.

Na slici 5, prikazan je rad ovog algoritma pod pretpostavkom da svaki objekat ima različitu vrednost (vrednosti atributa su označene kružićima, a tačke reza vertikalnim linijama).

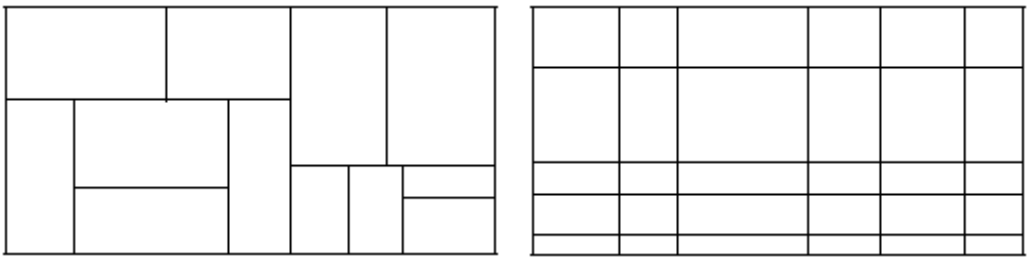


Slika 5. Equal frequency algoritam

3.2 OSNOVNE PODELE DISKRETIZACIJA

Postoji nekoliko podela na osnovu kojih je moguće klasifikovati algoritme za diskretizaciju. Prema [Ramırez-Gallego i dr., 2016] [Bay, 2015] neke od osnovnih podela su sledeće:

- Lokalna – Globalna diskretizacija (*local – global*). U slučaju kada se diskretizacija izvršava za pojedini atribut nezavisno od uticaja ostalih atributa, ona je lokalna [Nguyen, 2006]. Za tabelarno organizovane podatke podela po atributima (kolone) i objektima (vrste) se može prikazati grafički. Na slici 6 (slika je preuzeta iz [Nguyen, 2006]) grafički je prikazan primer lokalne i globalne diskretizacije.



Slika 6. Lokalna (levo) i globalna (desno) diskretizacija

- Dinamička – Statička diskretizacija (*dynamic – static*). Statički metod određuje maksimalan broj tačaka reza jednog atributa nezavisno od drugih. Dinamički metod upoređuje sve tačke reza za sve atribute istovremeno.
- Nadzirana – Nenadzirana diskretizacija (*supervised – unsupervised*). U slučaju kada se diskretizacija izvršava u zavisnosti od uticaja atributa odluke, ona je nadzirana. Inače je nenadzirana.
- Univarijantna – Multivarijantna diskretizacija (*univariate – multivariate*). U slučaju kada se analizira preklapanje tačaka reza čitavog skupa atributa, odnosno interakcija atributa, radi se o multivarijantnoj diskretizaciji [Muhlenbach, 2005].
- Deleća – Objedinjujuća diskretizacija (*splitting – merging, Top-Down - Bottom-up*). Ova podela je nastala na osnovu početno generisanih tačaka reza određenog algoritma. U slučaju ako neki algoritam diskretizacije za početni skup tačaka reza uzme tačke dobijene priemnom Naive algoritma, onda se konačan broj tačaka reza dobija objedinjavanjem početnih podintervala.
- Direktna – Inkrementalna diskretizacija (*direct – incremental*). Kod inkrementalane diskretizacije obično se posmatraju slojevi vrednosti atributa. Na osnovu originalnih vrednosti atributa nekom funkcijom se generiše novi sloj vrednosti atributa. Inkrementalna diskretizacija se naročito koristi kod stream podataka [Webb, 2014].
- Po meri ocene diskretizacije (informacija, statistika, grubi skupovi, wrapper [Ferreira i dr., 2014], ...).

Ako se u diskretizaciju uključi ili ne uključi ekspert, onda bi to bila još jedna podela. Stručnjak najbolje može da prilagodi tačke reza tako da odgovaraju važnosti određenog

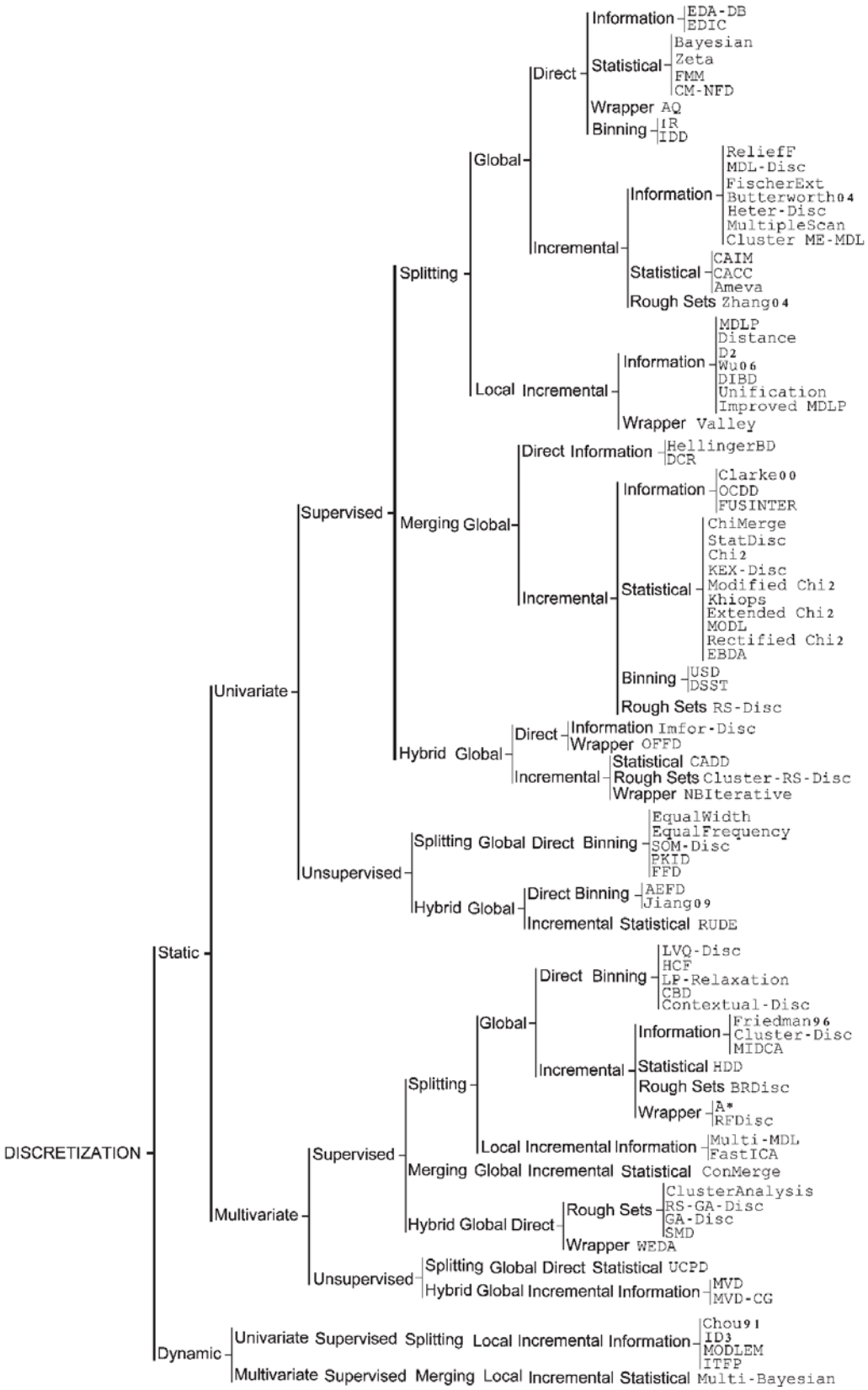
atributa. Međutim u situacijama kada ekspert ne razume „prirodu" određenog algoritma diskretizacije, to može da bude kontraproduktivno.

U radu [García i dr., 2013] istaknut je razvoj algoritama za diskretizaciju i dat je uporedni pregled algoritama diskretizacije preko grafa podele - slika 7 je preuzeta iz istog rada. Na osnovu podele algoritama diskretizacije rađena su komparativna ispitivanja algoritama diskretizacije i to uglavnom eksperimentalno na nekoliko izabranih baza. Entropija kao matematički alat je veoma korišćena u diskretizacijama. Komparirana je u odnosu na nadzirane i nenadzirane diskretizacije u okviru algoritma C4.5 [Dougherty i dr., 1995]. Posebno se mogu istaći komparacije algoritama diskretizacije baziranih na entropiji i algoritma diskretizacije na osnovu teorije grubih skupova [Nguyen, 1997; Düntsch i dr.,1998]. Zadnjih godina razvili su se algoritmi koji povezuju ideje ova dva algoritma kao što je primena entropije u diskretizaciji u okviru teorije grubih skupova [Shi, 2014], posebno sa čuvanjem konzistentnosti sistema [Cong, 2010] [Liu i dr., 2008].

Od svih podela diskretizacija iskristalisala se ideja o klasifikaciji na osnovu sledećih karakteristika: lokalna na prema globalnoj, statička na prema dinamičkoj i nadzirana prema nenadziranoj diskretizaciji.

Pored osnovnih podela analiziranje sledećeg svojstva algoritma diskretizacije je vredno pažnje: redukcija atributa. U tom slučaju se u samom početku rada algoritma diskretizacije omogućuje jednostavnije određivanje tačaka reza. Za redukciju atributa razvijane su razne heuristike, posebno u teoriji grubih skupova [Fayyad i dr., 1993] [Qian i dr., 2010 b].

Za poboljšanje rada algoritma diskretizacije u cilju generisanja kraćih pravila i klasifikacije u skladu sa tim, aproksimativna diskretizacija ima svoje mesto često kao metod nadgradnje određenog algoritma diskretizacije ili kao heuristika [Zhao i dr., 2009] [Ognjenovic i dr., 2011 a]. Razvojem aproksimativne diskretizacije dodatno je razvijan način predstavljanja bitne relacije određene teorije [Ognjenovic i dr., 2011 b].



Slika 7. Algoritmi diskretizacije na osnovu podela

3.3 DISKRETIZACIJA PODATAKA BAZIRANA NA ENTROPIJI

Entropija kao matematička formula za određivanje neodređenosti (neizvesnosti) se prema [Branović, 2003] koristi u sistema opisanim stanjima na sledeći način:

- neka sistem S ima različita stanja A_1, A_2, ... , A_n
- svako stanje A_i je opisano nekom verovatnoćom p_i (verovatnoćom da se desi stanje A_i)
- sistem S se onda može predstaviti slučajnom promeljivom opisanom sledećom formulom

$$S:\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad \sum_{i=1}^{n} p_i = 1 \tag{1}$$

- formula za entropiju kao neizvesnost izbora jedne od mogućih vrednosti kojima su pridružene verovatnoće p_i je prema [Shannon, 1951]:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2}$$

- u slučaju kada su verovatnoće p_i jednake, formula glasi

$$H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \log_2 n \tag{3}$$

Jedinica za entropiju je bit. Ako je raspodela verovatnoća P(0.5, 0.5) tada je H(S)=1; u slučaju raspodele P(0.67, 0.33) entropija je H(S)=0.92; za slučaj P(1, 0) dobija se H(S)=0.

Primer 3.3.1 (preuzet iz [Branović, 2003])

Odrediti entropiju fizičkog sistema koji se sastoji od dva aviona: lovca i bombardera, koji učestvuju u vazdušnom boju. Kao rezultat boja moguće je jedno od stanja sistema:

a. oba su nepogođena,
b. lovac pogođen, bombarder ne,
c. bombarder pogođen, lovac ne,
d. oba su pogođena

Verovatnoća da lovac bude pogođen je 0.4, a bombarder 0.5. Tada funkcija raspodele X izgleda:

$$S:\begin{pmatrix} a & b & c & d \\ 0.3 & 0.2 & 0.3 & 0.2 \end{pmatrix}$$

a entropija na osnovu formule (2):

$$H(S) = -2(0.2\log_2 0.2 + 0.3\log_2 0.3) = 2(0.4639 + 0.5211) = 1.97 bita$$

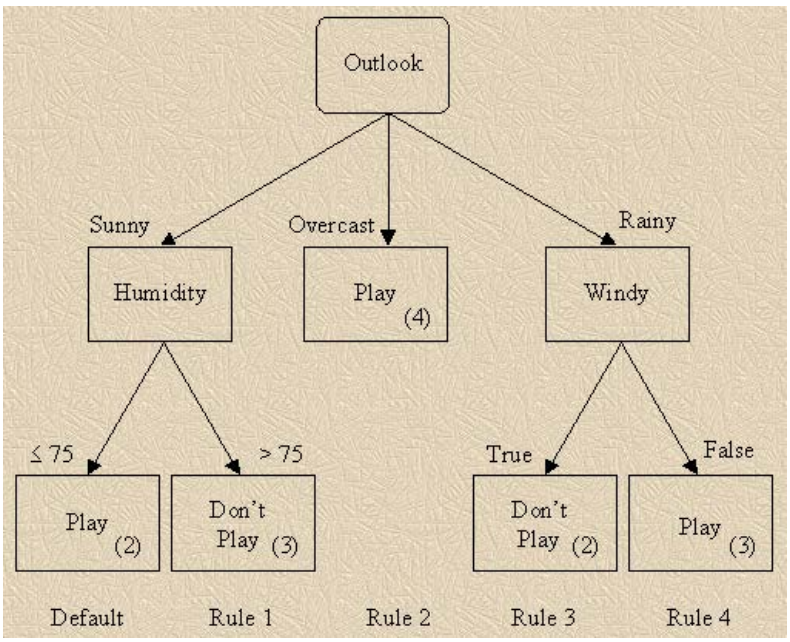
Diskretizacija podataka bazirana na entropiji svoj početak ima u razvoju algoritama za kreiranje stabala odlučivanja. Većina algoritama za kreiranje stabla odlučivanja su varijacije osnovnog algoritma ID3 koji je J. Ross Quinlan predstavio 1986. godine [Quinlan, 1993]. Postoje izuzeci koje je interesantno posmatrati u relaciji i sa stablima odlučivanja i sa teorijom grubih skupova [Brtka, 2008] [Ognjenovic i dr., 2012].

3.3.1 Algoritam ID3 i C4.5

Stabla odlučivanja su veoma korišćena metoda u predstavljanju tabelarnih podataka i grafički su nalik grafu tipa stablo. Na osnovu dobijenog grafa moguće je generisati IF ... THEN pravila koja opisuju podatke. Koren stabla predstavlja početak klasifikacije, a svaki drugi čvor neki atribut. Grana koja izlazi iz čvora određuje različite vrednosti za dati atribut. Krajnji čvor se zove list i njemu je pridružena odluka.

Primer 3.3.1.1. (preuzet sa [Quinlan, 2014])

Ovaj primer je Ross Quinlan postavio kao osnovni primer za stabla odlučivanja koja je definisao. Na osnovu određenih čvorova prikazanih na slici 8, generisana su pravila odlučivanja prikazana na slici 9.



Slika 8. Primer stabla odlučivanja

Rule 1 suggests that if "outlook = sunny" and "humidity > 75" then "Don't Play".
Rule 2 suggests that if "outlook = overcast" then "Play".
Rule 3 suggests that if "outlook = rain" and "windy = true" then "Don't Play".
Rule 4 suggests that if "outlook = rain" and "windy = false" then "Play".
Otherwise, "Play" is the default class.

Slika 9. IF THEN pravila dobijena sa stabla sa slike 8

Prvi korak u kreiranju stabla odlučivanja je određivanja atributa koji će biti u korenu stabla i čije grane će izvršiti podelu svih objekata. Algoritam ID3 za odabir atributa koristi podatak za informacionu dobit (information gain) na osnovu formule (4). Na osnovu [Milikić, 2016] informaciona dobit Gain(A,S) atributa A nad skupom instanci S predstavlja količinu informacija koja se dobija poznavanjem vrednosti atributa A, dok u odnosu prema entropiji predstavlja razliku entropije pre grananja i entropije nakon grananja nad atributom A. Zbog toga se formula (4) drugačije može zapisati i kao (5). Time informaciona dobit pokazuje koliko se entropija sistema smanjuje, ako se za odlučivanje koristi određen atribut, tj. koji atribut nosi najviše informacija [Runić, 2016]: informaciona dobit jednaka je razlici entropije čvora roditelja i prosečne entropije čvorova dece.

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} * H(S_j) \tag{4}$$

gde je: $H(S)$ - entropija celog skupa instance S; $|S_j|$ - broj instanci sa j-tom vrednošću atributa A; $|S|$ - ukupan broj instanci u skupu S; v - skup vrednosti atributa A; $H(S_j)$ - entropija podskupa instanci sa atributom A; $H(A,S)$ - entropija atributa A

$$Gain(A,S) = H(S) - H(A,S) \tag{5}$$

Algoritam C4.5 je proširenje ID3 algoritma, nastao optimizacijom generisanja delova stabla, u radu sa kontinualnim vrednostima atributa, nedostajućim vrednostima i drugim faktorima bitnim za poboljšavanje efikasnosti algoritma [C4.5 Tutorial, 2016]. Sledećim primerom pokazaće se generisanje stabla odlučivanja na osnovu entropije, odnosno informacione dobiti.

Primer 3.3.1.1. (preuzet iz [Wei, 2003])

Na osnovu tabele odlučivanja date tabelom I, na osnovu entropije treba da se generišu pravila odlučivanja.

No.	*A*	*B*	*C*	*D*	*d*
1	1	2	2	1	1
2	1	2	3	2	1
3	1	2	2	3	1
4	2	2	2	1	1
5	2	3	2	2	2
6	1	3	2	1	1
7	1	2	3	1	2
8	2	3	1	2	1
9	1	2	2	2	1
10	1	1	3	2	1
11	2	1	2	2	2
12	1	1	2	3	1

TABELA I. TABELA ODLUČIVANJA

Na osnovu formula (2) i (4), i vrednosti uslovnih atributa tabele I (A, B, C i D) i vrednosti atributa odluke d, može se izračunati da je:

$$H(S) = 0.811 bits$$

$$H(A,S) = 0.696 bits$$

$$H(B,S) = 0.784 bits$$

$$H(C,S) = 0.771 bits$$

$$H(D,S) = 0.729 bits$$

Na osnovu toga odgovarajuće informacione dobiti za atribute A, B, C i D prema formuli (5) su:

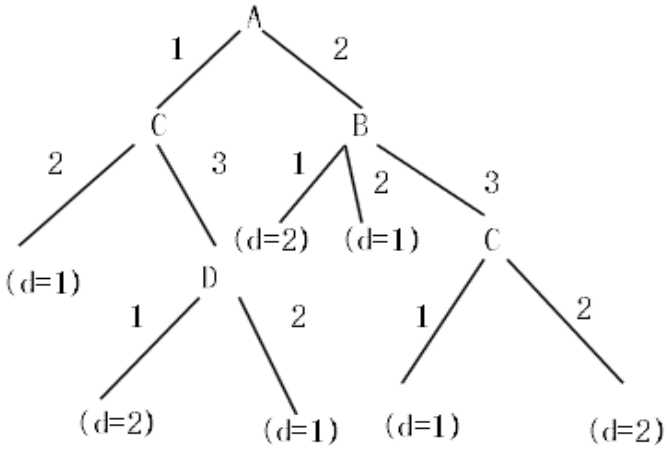
$$Gain(A,S) = 0.115 bits$$

$$Gain(B,S) = 0.027 bits$$

$$Gain(C,S) = 0.04 bits$$

$$Gain(D,S) = 0.082 bits$$

Time se zaključuje da će se odabirom čvora A kao korenog čvora stabla dobiti najviše informacija iz tabele. Čvor A se postavlja kao koreni čvor, a njegove dve vrednosti („1” i „2”) početnu tabelu dele na dva podskupa. Iteracije se ponavljaju nad preostalim podacima. Ako se na isti način izračunaju vrednosti za preostale delove tabele, stablo koje se dobije je prikazano na slici 10.



Slika 10. Stablo na osnovu entropije

Na osnovu prikazanog stabla dobilo se sedam IF THEN pravila:

1. IF A=1 AND C=2 THEN d=1
2. IF A=1 AND C=3 AND D=1 THEN d=2
3. IF A=1 AND C=3 AND D=2 THEN d=1
4. IF A=2 AND B=1 THEN d=2
5. IF A=2 AND B=2 THEN d=1
6. IF A=2 AND B=3 AND C=1 THEN d=1
7. IF A=2 AND B=3 AND C=2 THEN d=2

Dobijena pravila mogu biti različite dužine (sa različitim brojem uslovnih atributa) i sa samo jednom vrednosti odluke.

3.3.2 Algoritam diskretizacije baziran na entropiji

Metod diskretizacije atributa sa kontinualnim vrednostima baziran na entropiji je u radovima [Catlett, 1991] [Fayyad i Irani, 1993] opisan na osnovu heuristike minimalne entropije. U odnosu na entropiju kandidata particije atributa odluke određuju se oblasti za diskretizaciju. Za dati skup instanci S, atribut A i tačke T - particije oblasti, entropija klase po tački podele T se označava sa H(A, T; S) i računa na osnovu formule (6).

$$E(A,T;S)=\frac{|S_1|}{|S|}H(S_1)+\frac{|S_2|}{|S|}H(S_2) \tag{6}$$

Za dati atribut A oblast T_{min} koja minimizuje entropiju u odnosu na sve moguće particije (podele) oblasti se određuje kao oblast binarne diskretizacije. Postupak se primenjuje rekurzivno i zaustavlja na osnovu nekog uslova. Na osnovu [Fayyad i Irani, 1993] princip dužina minimalnih karakteristika (*minimum description length* - MDL), uslov koji zaustavlja proces daljeg deljenja je dat formulom (7).

$$Gain(A,T;S)<\frac{\log_2(N-1)}{N}+\frac{\Delta(A,T;S)}{N} \tag{7}$$

gde je

- N broj instanci u skupu S,
- $Gain(A,T;S)=H(S)-E(A,T;S)$,
- $\Delta(A,T;S)=\log_2(3^k-2)-[k\cdot H(S)-k_1\cdot H(S_1)-k_2\cdot H(S_2)]$, i
- k_i je broj podklasa iz skupa S_i.

S obzirom da se particije po svim grananjima rekurzivne diskretizacije procenjuju (po istim kriterijuma) nezavisno, neke oblasti kontinualnih vrednosti će biti podeljene veoma fino (sa malom entropijom), dok će druge biti podeljene grubo [Dougherty i dr.,

1995]. U okviru ove disertacije ova konstatacija će se jasno videti kao rezultat eksperimenta.

U odnosu na varijante osnovnog algoritma diskretizacije baziranog na entropiji, u odnosu na pridružene heuristike, moguća je disproporcija u ukupnom broju početnih kandidata tačaka reza, a samim tim i disproporcija u odnosu na brzinu rada algoritma [Fayyad i Irani, 1992].

Kao jedna varijanta algoritma diskretizacije baziranog na entropiji vredi pomenuti algoritam C4.5-Disc koji koristi algoritam C4.5 za diskretizaciju. Za razliku od algoritma opisanog u [Fayyad i Irani, 1993] koji koristi princip dužina minimalnih karakteristika za zaustavljanje rada, algoritam C4.5-Disc u toku grananja gradi kompletno stablo. Stablo se orezuje tako što se odstranjuju podstabla koja u maloj meri doprinose tačnosti klasifikacije. Podstabla koja imaju grešku klasifikacije veću od greške bez njega se odsecaju.

3.4 DISKRETIZACIJA PODATAKA NA OSNOVU MAKSIMALNE RAZBERIVOSTI

3.4.1 Teorija grubih skupova

Teoriju grubih skupova je razvio Pawlak 1982 za analizu podataka [Pawlak, 1982]. Osnovna namena grubih skupova je aproksimacija nepoznatih znanja preko poznatog [Skowron i dr., 1999]. Za teoriju grubih skupova je bitno postojanje univerzuma koji sadrži objekte definisane pomoću vrednosti svojih atributa. Bazirana na principu relacije nerazberivosti objekata i konceptu aproksimacije, ova teorija omogućuje prepoznavanje zavisnosti između atributa odluke i uslovnih atributa [Øhrn i dr., 1998]. Pored relacije nerazberivosti u teoriji grubih skupova moguće je raditi i sa relacijama sličnosti, tolerantnosti i drugim [Brtka i dr., 2011a].

Podaci koji se analiziraju su tabelarno organizovani. U teoriji grubih skupova definisana je informaciona tabela [Nguyen, 2006]. Informacionu tabelu čini uređena četvorka: $S = \langle U, Q, V, f \rangle$, gde je U konačan skup objekata – univerzum; $Q = \{q_1, q_2, ..., q_m\}$ je konačan skup atributa; $V = \bigcup_{q \in Q} V_q$, gde je V_q domen atributa q (vrednosti atributa); $f = U \times Q \to V$ je totalna funkcija takva da je $f(x,q) \in V_q$ za svako $q \in Q, x \in U$ i zove se informaciona funkcija (*information function*). Svaki objekat $x \in U$ je opisan vektorom:

$$\inf_q(x) = [f(x,q_1), f(x,q_2), ..., f(x,q_m)] \tag{8}$$

koji definiše vrednosti atributa objekta *x*.

Relacija nerazberivosti

Neka je sa *P* označen neprazan podskup skupa atributa *Q*. Nad objektima iz univerzuma *U* definisana je relacija I_P na sledeći način:

$$I_P = \{(x,y) \in U \times U : f(x,q) = f(y,q), \forall q \in P\} \tag{9}$$

Relacija (9) se zove relacija nerazberivosti, ili relacija nerazlikovanja (*indiscernibility relation*). Koristi se i oznaka *IND(P)*. Ako $(x,y) \in I_P$, kaže se da su objekti *x* i *y* P-nerazberivi (*P-indiscernible*). Relacija nerazberivosti je relacija ekvivalencije. Ovakva relacija generiše klase ekvivalencije. Familija klasa ekvivalencije koju generiše I_P označena je sa U/I_P. Klase ekvivalencije generisane relacijom I_P nazivaju se P-elementarni skupovi (*P-elementary sets*), a klasa ekvivalencije koja sadži objekat $x \in U$ označena je sa $I_P(x)$ ili $[x]_P$. Ako je $P = Q$, P-elementarni skupovi se nazivaju atomi (*atoms*).

Neka je *S* informaciona tabela, *X* neprazan podskup od *U* (kaže se grubi skup *X*), a $\varnothing \neq P \subseteq Q$. Tada je:

$$\underline{P}(X) = \{x \in U : I_P(x) \subseteq X\} \tag{10}$$

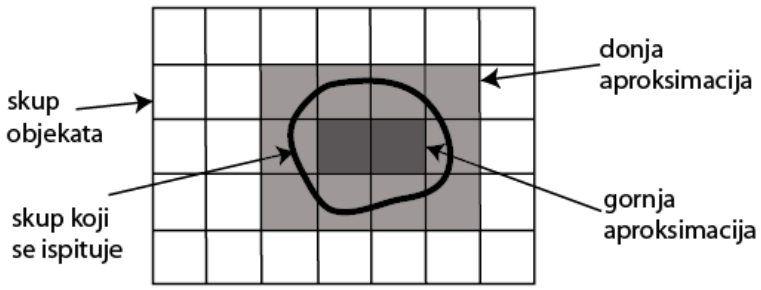
$$\overline{P}(X) = \bigcup_{x \in X} I_P(x) \tag{11}$$

Sa $\underline{P}(X)$ je označena P-donja aproksimacija (*P-lower approximation*), a sa $\overline{P}(X)$ P-gornja aproksimacija (*P-upper approximation*) skupa *X*. Elementi $\underline{P}(X)$ su oni objekti $x \in X$ koji pripadaju klasi ekvivalencije generisanoj sa I_P koja je sadržana u *X*. Elementi $\overline{P}(X)$ su oni objekti $x \in X$ koji pripadaju klasi ekvivalencije generisanoj sa I_P koja sadrži najmanje jedan objekat *x* koji pripada *X* [Pawlak, 1991] [Virginia, 2013].

P-granica (*P-boundary*) *X* u *S* definiše se kao:

$$Bn_P(X) = \overline{P}(X) - \underline{P}(X) \tag{12}$$

Grafička interpretacija P-granice je prikazana na slici 11. Kvadrati obojeni tamno sivom bojom predstavljaju klase objekata čiji svi elementi pripadaju skupu X (čitava klasa pripada skupu X) i time čine donju aproksimaciju grubog skupa X. Kvadrati obojeni svetlo sivom bojom predstavljaju klase objekata koje sadrže neke elemente koji pripadaju skupu X (sadrže i one elemente koji ne pripadaju skupu X) - takve klase predstavljaju granični region grubog skupa X. Elementi opisani istim osobinama klase iz graničnog regiona mogu ali ne moraju pripadati skupu X; za elemente graničnog regiona se ne može sa sigurnošću tvrditi da pripadaju skupu X – kaže se da su oni elementi aproksimacije skupa X [Brtka, 2008].



Slika 11. Osnovna ideja teorije grubih skupova

Ako je skup atributa *Q* informacione tabele podeljen na uslovne (*condition*) atribute $C \neq \varnothing$ i atribute odluke (*decision attributes*) $D \neq \varnothing$, tako da je $C \cup D = Q$ i $C \cap D = \varnothing$, takva informaciona tabela nazvana je tabela odluke (*decision table*). Atributi odluke *D*, generišu particiju skupa *U* preko relacije nerazberivosti I_D. D-elementarni skupovi se nazivaju klase odluke (*decision classes*). Tabela odluke predstavljena je uređenom četvorkom $S = \langle U, (C \cup D), V, f \rangle$.

U tabeli odluke S, klasa ekvivalencije $[x_i]_q$ objekta x_i definisana je kao skup svih objekata iz U koji su u relaciji nerazberivosti sa x_i u odnosu na atribut q (formula (13)).

$$[x_i]_q = \{x_j \in U \mid x_i I_q x_j\} \tag{13}$$

Konzistentnost

Generalizovana funkcija odluke (*generalized decision function*) $\partial_P(x_i)$ objekta x_i za skup $P \subseteq C$, definisana je formulom (14) kao skup klasa odluke po svim objektima u okviru klase ekvivalencije x_i [Chebrolu i Sanjeevi, 2015].

$$\partial_P(x_i) = \{ f(x_i, d) \mid x_i \in [x_i]_P \} \tag{14}$$

ili drugačije formulisano

$$\partial_P(x_i) = \{ i : \exists_{x' \in U} [(x' IND(P) x) \wedge (f(x', d) = i)] \}$$

gde $d \in D$.

Za tabelu odluke se kaže da je konzistentna (*consistent*) ako je kardinalnost od $\partial_P(x_i)$ jednaka 1 za sve objekte u univerzumu. Inače ako kardinalnost generalizovane funkcije odluke nije jednaka 1, tabela je inkonzistenta (*inconsistent*). Konzistentnost tabele je bitna osobina koja je u nekim drugim teorijama klasifikacije zanemarena.

Relacija razberivosti

Relacija nerazberivosti (I_P ili *IND(P)*) je relacija ekvivalencije čije klase se koriste u definisanju donje (*lower*) i gornje (*upper*) aproksimacije nekog skupa, kao što je prikazano u formulama (9), (10) i (11).

Komplement relacije nerazberivosti se zove P-razberiva relacija (*P – discernibility relation*), takođe koristi informacionu funkciju (8), a označava se sa:

$$\begin{aligned} DISC_S(P) &= U \times U - IND_s(P) \\ &= \{(x, y) \in U \times U : \inf_P(x) \neq \inf_P(y)\} \\ &= \{(x, y) \in U \times U : \exists_{q \in P} f(x, q) \neq f(y, q)\} \end{aligned} \tag{15}$$

Relacija $DISC_S(P)$ je monotona, odnosno za svaki $B, C \subset Q$ važi $B \subset C \Rightarrow DISC_S(B) \subset DISC_S(C)$.

Redukt

U teoriji grubih skupova informacioni redukt ili kraće redukt (*information reduct*, ili samo *reduct*) se intuitivno prepoznaje kao minimalni podskup atributa koji čuvaju razberivost između objekata [Brtka i dr., 2012]. Na osnovu definicije redukta date u [Nguyen, 2006], bilo koji podskup P od skupa atributa Q takav da je $DISC_S(P) = DISC_S(Q)$ se zove redukt od S. Skup svih redukta informacionog sistema S se označava sa $RED(S)$.

Ako se posmatra tabela odlučivanja, značajna je mogućnost opisivanja klasa odlučivanja pomoću podskupa uslovnih atributa. Ova mogućnost se može izraziti pomoću generalizovane funkcije odlučivanja date formulom (14).

Skup atributa $P \subseteq Q$ se zove relativan redukt (*relative reduct*) ili redukt odlučivanja (*decision reduct*) tabele odlučivanja S akko:

- $\partial_P(x) = \partial_Q(x)$ za sve objekte $x \in U$, i
- bilo koji pravi podskup od P ne zadovoljava prethodni uslov,

odnosno P je minimalan podskup skupa atributa koji zadovoljavaju da $\forall_{x \in U} \partial_P(x) = \partial_Q(x)$.

Veza između relacije razberivosti i redukta se na osnovu [Nguyen, 2006] može prikazati pomoću funkcije razberivosti:

- Informacioni redukti su tačno oni redukti koji odgovaraju funkciji razberivosti. Za proizvoljan skup atributa $P \subset Q$, broj parova objekata koji su razberivi atributima iz P je:

$$disc(P) = \frac{1}{2} card\left(DISC_S(P)\right)$$

- Relativni redukti tabele odlučivanja sa atributima $C \cup D$ su tačno oni redukti koji odgovaraju relativnoj funkciji razberivosti:

$$disc_{dec}(P) = \frac{1}{2} card\left(DISC_S(P) \cap DISC_S(\{dec\})\right)$$

za $dec \in D$. Relativna funkcija razberivosti vraća broj parova objekata iz različitih klasa odluke, koji su razberivi atributima iz P.

Primer 3.4.1.1 (primer je preuzet iz [Nguyen, 2006],)

U ovom primeru je analiziran problem „da li će se igra održati" u zavisnosti od vremenskih uslova. Objekti su opisani sa četiri uslovna atributa i atributom odluke sačinjenim od dve vrednosti. Posmatraće se prvih 12 objekata iz tabele II (preuzete iz [Nguyen, 2006]), gde je U = {1, 2, . . . , 12}, A = {a_1, a_2, a_3, a_4}, CLASSno = {1, 2, 6, 8}, CLASSyes = {3, 4, 5, 7, 9, 10, 11, 12}.

date	outlook	temperature	humidity	windy	play
ID	a_1	a_2	a_3	a_4	dec
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

TABELA II. TABELA ODLUČIVANJA „DA LI ĆE SE IGRA ODRŽATI”

Klase ekvivalencije relacije nerazberivosti $IND_s(P)$ za neke skupove atributa su date u tabeli III.

Skup atributa P	Klase ekvivalnecije relacije $IND_S(P)$
P={a_1}	{1, 2, 8, 9, 11}, {3, 7, 12}, {4, 5, 6, 10}
P={a_1, a_2}	{1, 2}, {4, 10}, {5, 6}, {7}, {8, 11}, {9}, {12}
P=Q={a_1, a_2, a_3, a_4}	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}, {11}, {12}

TABELA III. RELACIJA NERAZBERIVOSTI $IND_S(P)$ ZA NEKE SKUPOVE ATRIBUTA

Vrednosti relativne funkcije razberivosti $disc_{dec}(P)$ za sve podskupove $B \subset A$ su date u tabeli IV. Primećuje se postojanje dva relativna redukta $R_1 = \{a_1, a_2, a_4\}$ i $R_2 = \{a_1, a_3, a_4\}$.

Skupovi atributa P	$disc_{dec}(P)$
P={}	0
P={a_1}	23
P={a_2}	23
P={a_3}	18
P={a_4}	16
P={a_1, a_2}	30
P={a_1, a_3}	31
P={a_1, a_4}	29
P={a_2, a_3}	27
P={a_2, a_4}	28
P={a_3, a_4}	25
P={a_1, a_2, a_3}	31
P={a_1, a_2, a_4}	**32**
P={a_1, a_3, a_4}	**32**
P={a_2, a_3, a_4}	29
P={a_1, a_2, a_3, a_4}	**32**

TABELA IV. $disc_{dec}(P)$ ZA RAZLIČITE SKUPOVE

Problem dobijanja redukta je veoma složen [Nguyen i Slezak, 1999]. [Slezak, 2000]. Dobijanje minimalnog ili najkraćeg redukta (redukta minimalne kardinalnosti) je problem koji je NP težine [Skowron i Rauszer, 1992]. Dobijanje svih redukta je problem koji je najmanje NP težine.

Matrica razberivosti

Za tabelu odlučivanja $S = \langle U, (C \cup D), V, f \rangle$ gde je skup objekata $U = \{u_1, u_2, \ldots, u_n\}$ i skup uslovnih atributa $C = \{c_1, c_2, \ldots, c_k\}$, matrica razberivosti (*discernibility matrix*) je matrica

$$M(S) = \left[M_{i,j}\right]_{ij=1}^{n}$$

gde je $M_{i,j} \subset C$ skup atributa razberivih po objektima u_i i u_j, odnosno

$$M_{i,j} = \{c_m \in C : c_m(u_i) \neq c_m(u_j)\}$$

Skup Bulovih promenljivih $\{x_1, \ldots, x_k\}$ se pridružuje atributima $c_1, c_2, \ldots, c_k$ tako da za neki podskup $B \subset C$, $X(B)$ označava skup Bulovih promenljivih koje su pridružene atributima iz B. Time se problem dobijanja redukta prevodi u problem traženja odgovarajućih promenljivih. Za dva objekta $u_i, u_j \in U$, Bulova klauza χ_{u_i,u_j} koja se zove razberiva klauza (*discernibility clause*) se definiše na sledeći način:

$$\chi_{u_i,u_j}(x_1,\ldots,x_k)=\begin{cases}\sum_{c_m\in M_{i,j}} x_m & if\ M_{i,j}\neq 0\\ 1 & if\ M_{i,j}=0\end{cases}$$

Funkcija razberivosti

U cilju generisanja redukta pravi se Bulova funkcija f_S tako da je skup atributa redukt od S akko on odgovara prostim implikantama od funkcije f_S. Takva Bulova funkcija se može definisati na sledeći način:

1. Za (informacioni) redukt funkcija je oblika:

$$f_S(x_1,\ldots,x_k)=\prod_{i\neq j}\left(\chi_{u_i,u_j}(x_1,\ldots,x_k)\right) \tag{16}$$

   gde je χ_{u_i,u_j} razberiva klauza. Funkcija (16) je definisana u radu [Skowron i Rauszer, 1992] i zove se funkcija razberivosti (*discernibility function*).

2. Za relativni redukt funkcija je oblika:

$$f_S^{dec}(x_1,\ldots,x_k)=\prod_{i,j:dec(u_i)\neq dec(u_j)}\left(\chi_{u_i,u_j}(x_1,\ldots,x_k)\right) \tag{17}$$

Funkcija f_S^{dec} se zove funkcija razberivosti u odnosu na atribut odluke (*decision oriented discernibility function*).

Pravila odlučivanja

Kao i u drugim teorijama za klasifikaciju, pravila odlučivanja u teoriji grubih skupova su logičke formule koje pokazuju vezu između uslovnih atributa i atributa odluke. Prema [Bazan i dr., 2000], za tabelu odlučivanja S, implikacija $\alpha\Rightarrow\beta$, gde je α Bulov izraz od uslovnih atributa (na osnovu konjukcije, disjunkcije i negacije) i β Bulov izraz od atributa odluke (na osnovu konjukcije, disjunkcije i negacije) je pravilo odlučivanja u S.

U skladu sa korišćenim oznakama, svaki objekat iz tabele odlučivanja se može zapisati u obliku sledećeg pravila odlučivanja:

$$\bigwedge_{c_i\in C} c_i(x)\Rightarrow dec(x)$$

Opštije, bilo koja implikacija

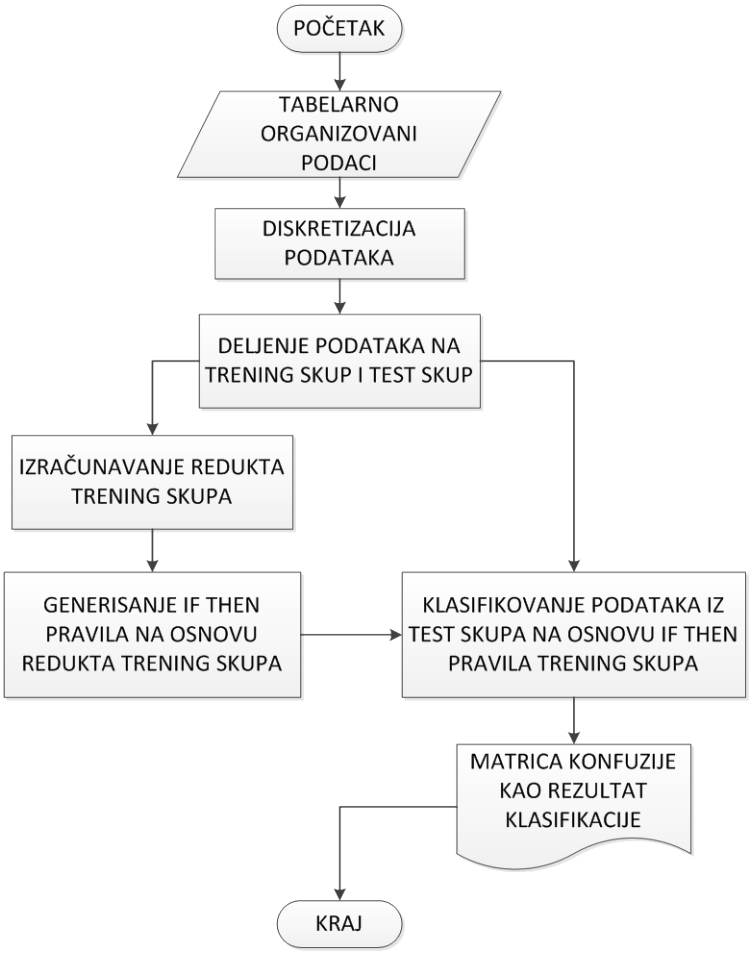
$$(c_i=v_1)\wedge\ldots\wedge(c_{i_m}=v_m)\Rightarrow(dec=k)$$

gde $c_{i_j} \in C$ i $v_j \in V_{c_{i_j}}$ zove se pravilo odlučivanja za k-tu klasu odluke [Bazan i dr., 2003].

Pravila odlučivanja se mogu generisati na osnovu formula za P-gornju, P-donju aproksimaciju, a takođe i na osnovu redukta. U okviru Bulovskog rezonovanja postoje veze između funkcije razberivosti, redukta i minimalnih pravila odlučivanja [Sengupta i Das,.2014].

Klasifikacija

Za tabelarno organizovane podatke, klasifikacija na osnovu teorije grubih skupova se može raditi (uobičajeno je) na način opisan blok-shemom sa slike 12.



Slika 12. Klasifikacija podataka na osnovu teorije grubih skupova

Deljenje podataka na trening i test skup se radi proizvoljno, uglavnom od 50% : 50% do 70% : 30% u korist trening skupa.

3.4.2 Osnovne definicije diskretizacije u teoriji grubih skupova

Diskretizacija kontinualnih podataka u okviru teorije grubih skupova je bazirana na definisanju skupa tačaka reza (*set of cuts*) nad svim atributima sa kontinualnim vrednostima. Neka je V_c skup vrednosti atributa $c \in C$. Neka je l_c leva granica a r_c desna granica skupa V_c tako da je $l_c < r_c$. Skup $V_c = [l_c, r_c) \subset R$, gde je R skup realnih brojeva. Neka je p_i realan broj takav da je $l_c \leq p_i < r$. Broj p_i pravi particiju objekata univerzuma *U* na dva disjunktna skupa U_l i U_r gde je

$$U_l = \{x_j \in U \mid f(x_j, c) \leq p_i$$

i

$$U_r = \{x_j \in U \mid f(x_j, c) > p_i\}$$

Oba skupa U_l i U_r su neprazna. Realan broj p_i definiše se kao tačka reza (*cut*) atributa *c*. Neka je P_c skup tačaka reza atributa c definisan sa $P_c = \{p_1, p_2, ..., p_k\}$, tako da je $l_c \leq p_1 < p_2 < ... < p_k < r_c$.

Neka je skup vrednosti atributa *c* definisan sa $V_c = \{v_1, v_2, \ldots, v_{|V_c|}\}$ tako da je $v_1 < v_2 < \ldots < v_{|V_c|}$. Skup osnovnih (bazičnih) tačaka reza B_c atributa *c* je definisan formulom (18).

$$B_c = \left\{ \frac{(v_1 + v_{2,})}{2}, \frac{(v_2 + v_{3,})}{2}, \ldots, \frac{(v_{|V_c|-1} + v_{|V_c|})}{2} \right\} \tag{18}$$

Skup osnovnih tačaka reza B skupa uslovnih atributa C je definisan kao:

$$B = \bigcup_{c \in C} B_c \tag{19}$$

Prema [Chebrolu i Sanjeevi, 2015] diskretizovana verzija konzistentnog sistema S je nov sistem odluke P-diskretizacija od S i on je definisan kao petorka $S^p = \langle U, (C \cup D), V, P, f^p \rangle$, gde je P skup tačaka reza (cuts) nad C, što se može zapisati kao

$$P = \bigcup_{c \in C} P_c \tag{20}$$

a funkcija f^P je definisana na sledeći način:

$$f^p(x_j, c) = \begin{cases} 0, & if \quad f(x_j, c) < p_1 \\ i, & if \ f(x_j, c) \in [p_i, p_{i+1}), 1 \leq i \leq k-1 \\ k, & if \quad f(x_j, c) \geq p_k \end{cases} \tag{21}$$

Ako su generalizovane funkcije $\partial_C(S)$ i $\partial_C(S^P)$ odgovarajućih tabela odlučivanja S i S^P iste, onda se za skup tačaka reza *P* kaže da je *S*-konzistentan (*S-consistent*). Ako za skup P važi da je *S*-konzistentan i da za $\forall P' \subset P \mid P'$ nije *S*-konzistentan, tada se za skup tačaka reza *P* kaže da je *S*-nesvodljiv (*S-irreducible*). Ako je skup P *S*-konzistentan i $\forall$ *S*-konzistentan skup tačaka reza P' važi da je $|P| \leq |P'|$, tada se skup tačaka reza *P* zove *S*-optimalan (*S-optimal*) u odnosu na kardinalnost skupa tačaka reza.

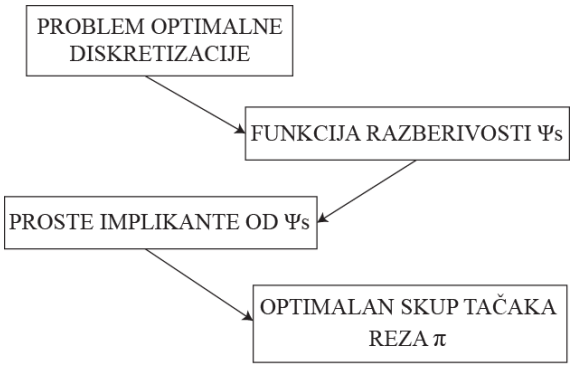
Problem pronalaženja *S*-optimalnog skupa tačaka reza je problem optimizacije [Chebrolu i Sanjeevi, 2015]. U teoriji grubih skupova se koristi Bulovski pristup (*Boolean reasoning approach*) za pronalaženje *S*-optimalnog skupa tačaka reza definisanjem matrice razberivosti i funkcije razberivosti [Jiang i dr., 2010] [Skowron i Rauszer, 1992].

3.4.3 Algoritam maksimalne razberivosti MD

Na osnovu skupa osnovnih tačaka reza iz formule (19), na osnovu [Nguyen, 2006] može se označiti da je skup:

$$P_{c_m} = \{\underbrace{(c_m, p_1^m)}_{p_1^{a_m}}, \quad \dots \quad \underbrace{(c_m, p_{n_m}^m)}_{p_{n_m}^{a_m}}\} \tag{22}$$

skup kandidata tačaka reza atributa $c_m \in C$. Algoritam maksimalne razberivosti na samom početku radi kodiranje opisano slikom 13 [Nguyen, 2006].



Slika 13. Bulovski pristup problemu optimalne diskretizacije

Svakoj tački reza $(c_m, p_i^m) \in P_{c_m}$ pridružuje se Bulova promenljiva p_i^m i pravi se skup $BCuts_{c_m} = \{p_1^{c_m}, \dots, p_{n_m}^{c_m}\}$ Bulovih promenljivih koji odgovara kandidatima tačaka reza atributa c_m. Za svaki skup tačaka reza $X \subset P$ sa $\sum_X$ se označava disjunkcija, odnosno sa $\prod_X$ se označava konjukcija Bulovih promenljivih koje odgovaraju tačkama reza iz skupa X [Nguyen i Nguyen, 1998].

Na osnovu formule (20) skup tačaka reza je $P = \bigcup_{c \in C} P_c$. Preciznije označeno $P = \bigcup P_{c^m}$. Za bilo koji par objekata $u_i, u_j \in U$ označava se skup tačaka reza $X_{i,j}^c$ iz P_c razberiv po u_i i u_j [Nguyen, 2006]:

$$X_{i,j}^c = \left\{ \left(c; p_k^c\right) \in P_c : \left(c\left(u_i\right) - p_k^c\right)\left(c\left(u_j\right) - p_k^c\right) < 0 \right\}$$

Neka je $X_{i,j} = \bigcup_{c \in C} X_{i,j}^c$. Funkcija razberivosti $\psi_{i,j}$ za par objekata u_i, u_j definiše se kao disjunkcija promenljivih koje odgovaraju tačkama reza iz $X_{i,j}$, odnosno

$$\psi_{i,j} = \begin{cases} \sum_{X_{i,j}} & ako\ X_{i,j} \neq \{\} \\ 1 & ako\ X_{i,j} = \{\} \end{cases} \tag{23}$$

Bulova funkcija razberivosti od S definiše se sa:

$$\Phi_S = \prod_{d\left(u_i\right) \neq d\left(u_j\right)} \psi_{i,j} \tag{24}$$

Postoji teorema (na osnovu [Nguyen, 1997]) koja glasi:

Teorema. Za bilo koji skup tačaka reza X:

1. X je S-konzistentan ako i samo ako je $\Phi_S\left(A_X\right) = 1$, gde je A_X karakteristična funkcija skupa definisanog formulom (19);
2. X je S-nesvodljiv ako i samo ako konjukcija $\prod_X$ je prosta implikanta od Φ_S;
3. X je S-optimalan ako i samo ako konjukcija $\prod_X$ je najkraća prosta implikanta funkcije Φ_S.

Datom teoremom se dobija da je problem traženja optimalnog skupa tačaka reza za tabelu odlučivanja polinomijalno svodljiv na problem traženja minimalne proste implikante monotone Bulove funkcije.

Primer 3.4.3.1 (primer je preuzet iz [Nguyen, 2006])

Neka tabela odlučivanja ima dva uslovna atributa a i b i sedam objekata $u_1, u_2, u_3, u_4, u_5, u_6, u_7$. Vrednosti atributa a i b, kao i vrednosti atributa odluke d date su u tabeli V (tabela je preuzeta iz [Nguyen, 2006]). Diskretizacija je urađena na osnovu skupa tačaka reza $C = \{(a;0.9), (a;1.5), (b;0.75), (b;1.5)\}$ dobijenog na osnovu izabrane četvrte proste implikante iz disjunktivne normalne forme formule razberivosti Φ_S.

$\mathbb{S}$	a	b	d
u_1	0.8	2	1
u_2	1	0.5	0
u_3	1.3	3	0
u_4	1.4	1	1
u_5	1.4	2	0
u_6	1.6	3	1
u_7	1.3	1	1

(a)

$\Longrightarrow$

$\mathbb{S}\|_C$	$a\|_C$	$b\|_C$	d
u_1	0	2	1
u_2	1	0	0
u_3	1	2	0
u_4	1	1	1
u_5	1	2	0
u_6	2	2	1
u_7	1	1	1

(b)

TABELA V. (A) ORIGINALNA TABELA ODLUČIVANJA S; (B) C-DISKRETIZACIJA OD S

Skup Bulovih promenljivih definisanih sa S je jednak $BCuts_S = \left\{p_1^a, p_2^a, p_3^a, p_4^a, p_1^b, p_2^b, p_3^b\right\}$, gde je $p_1^a = [0.8;1)$, $p_2^a = [1;1.3)$, $p_3^a = [1.3;1.4)$, $p_4^a = [1.4;1.6)$, $p_1^b = [0.5;1)$, $p_2^b = [1;2)$, $p_3^b = [2;3)$. Formule razberivosti $\psi_{i,j}$ za različite parove (u_i, u_j) objekata iz U (prema formuli 22) su sledeće:

$$\psi_{2,1} = p_1^a + p_1^b + p_2^b$$
$$\psi_{2,4} = p_2^a + p_3^a + p_1^b$$
$$\psi_{2,6} = p_2^a + p_3^a + p_4^a + p_1^b + p_2^b + p_3^b$$
$$\psi_{2,7} = p_2^a + p_1^b$$
$$\psi_{3,1} = p_1^a + p_2^a + p_3^b$$
$$\psi_{3,4} = p_2^a + p_2^b + p_3^b$$
$$\psi_{3,6} = p_3^a + p_4^a$$
$$\psi_{3,7} = p_2^b + p_3^b$$
$$\psi_{5,1} = p_1^a + p_2^a + p_3^a$$
$$\psi_{5,4} = p_2^b$$
$$\psi_{5,6} = p_4^a + p_3^b$$
$$\psi_{5,7} = p_3^a + p_2^b$$

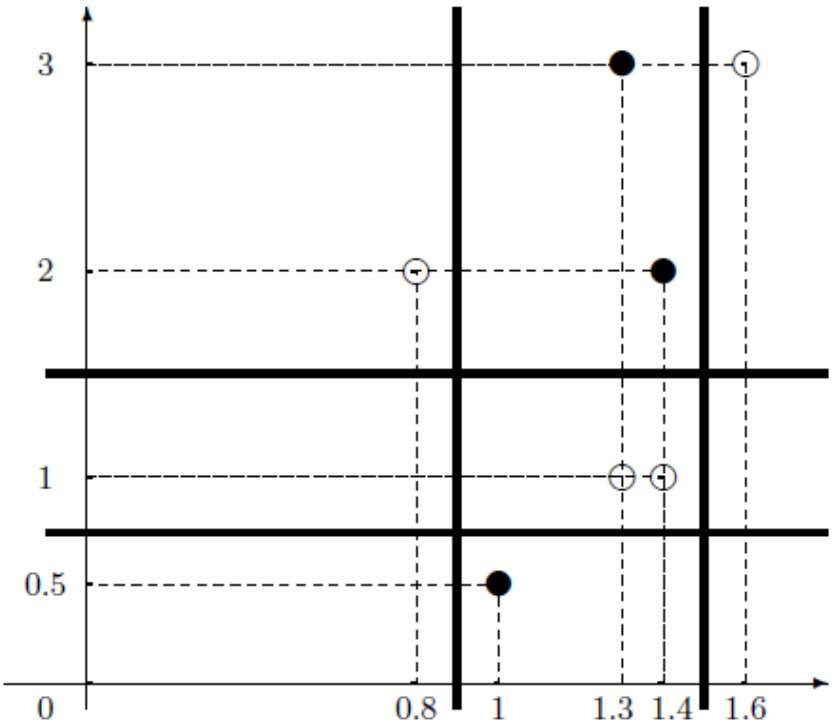
Formula Φ_S - Bulova funkcija razberivosti, prema formuli (24) u konjuktivnoj normalnoj formi je oblika:

$$\Phi_S = \left(p_1^a + p_1^b + p_2^b\right)\left(p_2^a + p_3^a + p_1^b\right)\left(p_2^a + p_3^a + p_4^a + p_1^b + p_2^b + p_3^b\right)\left(p_2^a + p_1^b\right)$$
$$\left(p_1^a + p_2^a + p_3^b\right)\left(p_2^a + p_2^b + p_3^b\right)\left(p_3^a + p_4^a\right)\left(p_2^b + p_3^b\right)\left(p_1^a + p_2^a + p_3^a\right)p_2^b$$
$$\left(p_4^a + p_3^b\right)\left(p_3^a + p_2^b\right)$$

Formula Φ_S transformisana u disjuktivnu normalnu formu je:

$$\Phi_S = p_2^a p_4^a p_2^b + p_2^a p_3^a p_2^b p_3^b + p_3^a p_1^b p_2^b p_3^b + p_1^a p_4^a p_1^b p_2^b$$

Ako se izabere četvrta prosta implikanta, dobije se skup tačaka reza $C = \{(a;0.9),(a;1.5),(b;0.75),(b;1.5)\}$. U tabeli 12 (b) je prikazana diskretizovana tabela na osnovu skupa tačaka reza C. Geometrijska interpretacija objekata i klasa odluke prikazana je na slici 14.



Slika 14. Geometrijsko predstavljanje podataka i tačaka reza

<u>Diskretizacija izračunavanjem redukta</u>

U radu [Nguyen, 1998] je pokazano da je problem diskretizacije za datu tabelu odlučivanja S polinomijalno ekvivalentan problemu izračunavanja redukta tabele odlučivanja S^* dobijene iz S, gde je $S^* = \{U^*, C^* \cup \{d^*\}\}$ i gde je definisano sledeće:

- $U^* = \{(u_i, u_j) \in U \times U : (i < j) \wedge (d(u_i) \neq d(u_j))\} \cup \{new\}$, gde $new \notin U \times U$, i predstavlja veštački element koji se koristi kao pomoć;
- $d^* : U^* \to \{0,1\}$ i definisana je sa $d^*(x) = \begin{cases} 0 \; ako \; x = new \\ 1 \qquad inace \end{cases}$
- $C^* = \{p_s^c : c \in C \quad s \; je \quad indeks \, od \quad [v_s^c, v_{s+1}^c) \; za \quad c\}$

Za bilo koji $p_s^c \in C^*$, vrednost $p_s^c\left(\left(u_i, u_j\right)\right)$ je jednaka 1 ako je:

$$\left[v_s^c, v_{s+1}^c\right) \subseteq \left[\min\left\{c\left(u_i\right), c\left(u_j\right)\right\}, \max\left\{c\left(u_i\right), c\left(u_j\right)\right\}\right)$$

a nula inače. Zbog toga se piše $p_s^c(new) = 0$. Sledeća propozicija je dokazana u [Nguyen, 1997])

Propozicija

Problem traženja nesvodljivog skupa tačaka reza je polinomijalno ekvivalentan problemu traženja relativnog redukta tabele odlučivanja.

MD heuristika

Problem optimalne diskretizacije može da se transformiše u problem minimalnog redukta [Nguyen, 2006]. U skladu sa tim svaka tačka reza se može pridružiti skupu parova objekata koji su razberivi tom tačkom reza. Zbog toga se optimalan skup tačaka reza može posmatrati kao minimalno pokrivanje skupa svih konfliktnih parova objekata, tj. objekata iz različitih klasa odluke.

Na osnovu toga, svaka tačka reza se može pridružiti parovima objekata koji su razberivi sa tom tačkom. Time se optimalan skup tačaka reza posmatra kao minimalan pokrivač skupa svih konfliktnih parova objekata (objekata iz različitih klasa odluke). {Za par objekata se kaže da su u konfliktu ako pripadaju različitim klasama odluke.} Algoritam maksimalne razberivosti MD heuristic je zbog toga pohlepni algoritam za minimalan skup pokrivanja. Baziran na strategiji najpre najbolji (best first search strategy) on uvek bira tačku reza koja razbire maksimalan broj konfliktnih parova objekata – ovaj korak se ponavlja sve dok su svi konfliktni parovi razberivi tom tačkom reza [Nguyen i Nguyen, 1996]. Slika pseudokoda algoritma MD-heuristics je data slikom 15, koja je preuzeta iz [Nguyen, 2006].

Algorithm 3. MD-heuristic for the optimal discretization problem

Input: Decision table $\mathbb{S} = (U, A, dec)$.
Output: The semi-optimal set of cuts.
begin
1 Construct the table $\mathbb{S}^*$ from $\mathbb{S}$ and set $\mathbf{B} := \mathbb{S}^*$;
2 Select the column of $\mathbf{B}$ with the maximal number of occurrences of 1's;
3 Delete from $\mathbf{B}$ the selected column in Step 2 together with all rows marked in this column by 1;
4 **if** $\mathbf{B}$ *consists of more than one row* **then**
Go to Step 2;
else
Return the set of selected cuts as a result;
Stop;
end
end

Slika 15. Slika pseudokoda MD-heuristics

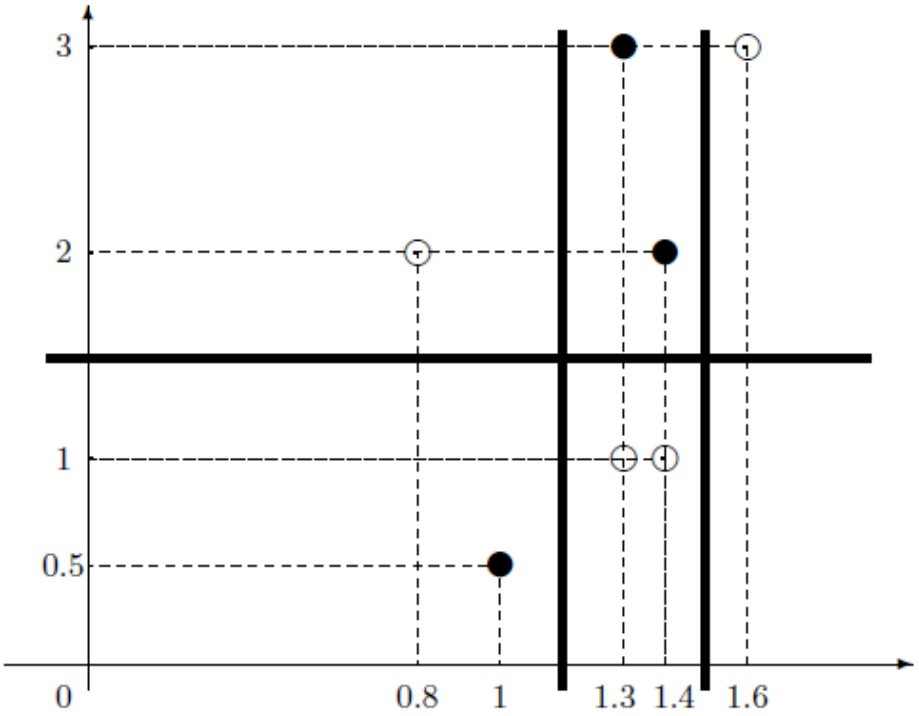
Na slici 16 je prikazan minimalan skup tačaka reza za tabelu odlučivanja V (a). Bez ulaženja u postupnu analizu generisanja tabele S^* (dobijene za tabelu odlučivanja V (a)), ona je prikazana u tabeli VI.

$\mathbb{S}^*$	p_1^a	p_2^a	p_3^a	p_4^a	p_1^b	p_2^b	p_3^b	d^*
(u_1, u_2)	1	0	0	0	1	1	0	1
(u_1, u_3)	1	1	0	0	0	0	1	1
(u_1, u_5)	1	1	1	0	0	0	0	1
(u_4, u_2)	0	1	1	0	1	0	0	1
(u_4, u_3)	0	0	1	0	0	1	1	1
(u_4, u_5)	0	0	0	0	0	1	0	1
(u_6, u_2)	0	1	1	1	1	1	1	1
(u_6, u_3)	0	0	1	1	0	0	0	1
(u_6, u_5)	0	0	0	1	0	0	1	1
(u_7, u_2)	0	1	0	0	1	0	0	1
(u_7, u_3)	0	0	0	0	0	1	1	1
(u_7, u_5)	0	0	1	0	0	1	0	1
new	0	0	0	0	0	0	0	0

TABELA VI. TABELA S^* DOBIJENA IZ TABELE S

Na osnovu rada algoritma MD, pronalaženje najkraće proste implikante nad tabelom S^* je rezultovalo prostom implikantom $p_2^a p_4^a p_2^b$ i odovarajućim minimalnim skupom tačaka reza $C = \{(a;1.15), (a;1.5), (b;1.5)\}$.

Na slici 15 je prikazana grafička interpretacija minimalanog skupa tačaka reza za tabelu odlučivanja V (a).



Slika 16. Minimalan skup tačaka reza

Veličina tabele S^* je $O(nk \cdot n^2)$ gde je n broj objekata a k broj kolona u S. Vremenska kompleksnost MD algoritma je prema [Nguyen, 2006] $O(n^3 k \times |C|)$, gde je C rezultujući skup tačaka reza. Algoritam maksimalne razberivosti je implementiran u dva vodeća sistema bazirana na teoriji grubih skupova koji se koriste u nauci i privredi: RSES [RSES, 2014] i Rosetta [Øhrn, 1998].

4 PREGLED STANJA U PODRUČJU ISTRAŽIVANJA

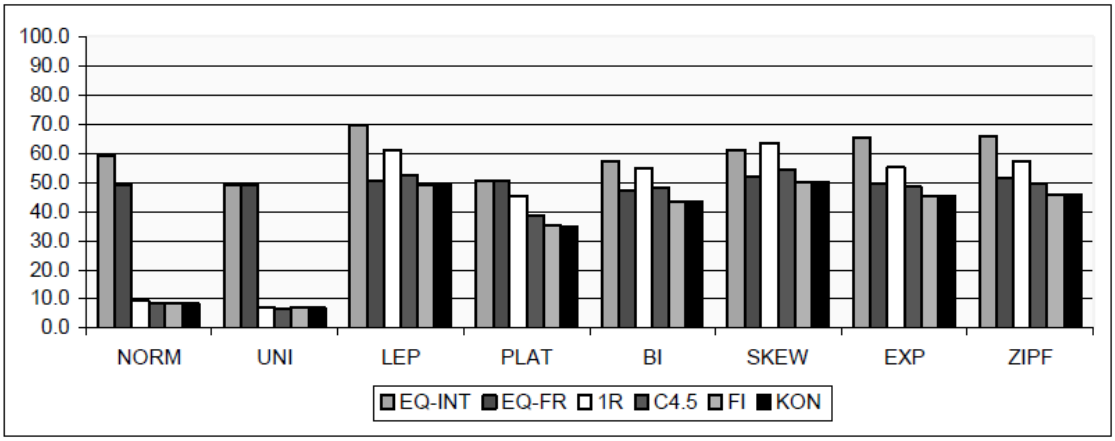
Postoji veliki broj radova u kojima je rađena komparacija raznih algoritama diskretizacije u odnosu na kvalitet klasifikacije – takva komparacija je rađena statistički nad određenim brojem baza [Dougherty i dr., 1995] [Bazan i dr., 2000] [Zhang i Cheung, 2014] [Nguyen i dr., 2014].

U odnosu na modifikaciju nekog konkretnog algoritma, komparacije su rađene bez osvrta na konzistentnost tabele odlučivanja. Empirijski rezultati pokazuju da kvalitet klasifikacijskih metoda zavisi od algoritma diskretizacije koji se koristi [Gama i dr., 2015].

4.1.1 Odnos raspodele podataka i algoritama diskretizacije

U radu [Ismail i Ciesielski, 2003] pokazan je uticaj raspodele podataka na šest popularnih diskretizacija. To je jedino istraživanje koje je primećeno iz direktne analize uticaja različitih raspodela podataka na konkretni algoritam diskretizacije. U njemu su obuhvaćene sledeće raspodele i algoritmi diskretizacije:

- Podaci su veštačko generisani tako da njihove raspodele pripadaju sledećim raspodelama: Normal, Uniform, Leptokurtic, Platykurtic, Bimodal, Skewed, Exponential i Zipf.
- Od algoritama diskretizacije korišćeni su:
    1. *Equal Interval Width*
    2. *Equal Frequency/*
    3. *Holte's 1R Discretizer*
    4. *C4.5 Discretizer*
    5. *Fayyad and Irani's Entropy Based MDL Method*
    6. *Kononenko's Entropy Based MDL Method*
- Pokazani su statistički odnosi vezani za kvalitet klasifikacije na osnovu relativne kvadratne greške i određene raspodele. Sumirane vrednosti su prikazane na slici 17. (slika preuzeta iz [Ismail i Ciesielski, 2003]).



Slika 17. Prosečna relativna kvadratna greška za sve raspodele i izabrane algoritme diskretizacije

U rada [Ismail i Ciesielski, 2003] nije uzet ni jedan algoritam diskretizacije baziran na konceptu teorije grubih skupova ili konceptu razberivosti. Od šest algoritama za diskretizaciju tri (*C4.5 Discretizer, Fayyad and Irani's Entropy Based MDL Method i Kononenko's Entropy Based MDL Method*) su bazirani na entropiji.

U okviru algoritma maksimalne razberivosti rađen je odnos raspodele podataka i velikih baza ali samo pomoću medijane [Nguyen, 2006], što isključuje analizu same funkcije raspodele.

Za tokove podataka, u radu [Gama i Pinto, 2006] predlaže se inkrementalni algoritam koji bi raspodelu podataka blago modifikovao kako bi se bolje primenjivao neki izabrani algoritam diskretizacije poput predloženog algoritma recursive entropy discretization.

Algoritam procene gustine histograma na osnovu MDL principa (*MDL Histogram Density Estimation*), iako se ne bavi direktnim analiziranjem histograma podataka, koristi entropiju za generisanje histograma pravilnosti podataka, čijom analizom se omogućuje diskretizacija [Kontkanen i Myllymaki, 2007]. Osnovna ideja je da se svaka pravilnost u podacima može koristiti za kompresiju podataka, tako da za postojanje veće mere pravilnosti, podaci mogu više da se kompresuju. Na osnovu principa MDL, učenje se izjednačuje sa pronalaženjem pravilnosti u podacima.

4.1.2 Smanjenje tačaka reza – aproksimativne dikretizacije

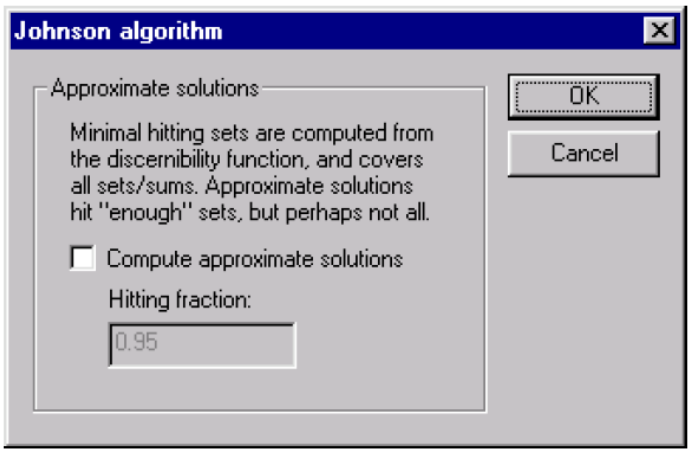
Optimizacija algoritama diskretizacije je često vezana za smanjenje broja atributa koji se diskretizuju [Chebrolu i Sanjeevi, 2015] [Slezak, 2002] [Qian i dr., 2010 b]. Smanjenjem broja atributa automatski se smanjuje broj kandidata tačaka reza. Međutim redukcija atributa ne omogućava uticaj redukovanog atributa na generisanje tačaka reza preostalih atributa.

Za pojam aproksimativna diskretizacija vezuju se različiti metodi. U odnosu na određenu formulu kojom se dobijaju tačke reza, aproksimativna diskretizacija može da predstavlja aproksimaciju izračunavanja vrednosti tačke reza [Patwardhan, 2016]. U slučaju primene neke teorije ili metoda aproksimacija može da se ogleda u pridruživanju određene heuristike koja je aproksimacija nekih znanja, u određivanju tačaka reza. Na osnovu preovlađujuće normalne raspodele podataka, napravljen je Approximate Equal Frequency metod za diskretizaciju [Sheng-yi i dr., 2009].

Aproksimacija diskretizacije u ovoj disertaciji posmatrana je sa stanovišta smanjenja tačaka reza u odnosu na izabrani algoritam diskretizacije. Smanjenje tačaka reza ne mora nužno značiti redukciju atributa, odnosno izbacivanje tačaka reza je moguće tako da ostane samo jedna tačka reza za dati atribut.

U okviru Teorije grubih skupova, prikazani algoritam maksimalne razberivosti MD kao inkrementalni algoritam određuje proste implikante Bulove funkcije koja je data formulom (24) dobijene nad svim kandidatima tačaka reza. Aproksimativno rešenje diskretizacije može se dobiti izračunavanjem aproksimativnih prostih implikanti Bulove funkcije [Øhrn i dr., 1998]. Ovo rezultuje manjim brojem tačaka reza na štetu povećanja inkonzistentnosti diskretizovane tabele odlučivanja [Øhrn, 2001]. U sistemu Rosetta moguće je izračunavanje aproksimativnog rešenja diskretizacije algoritma

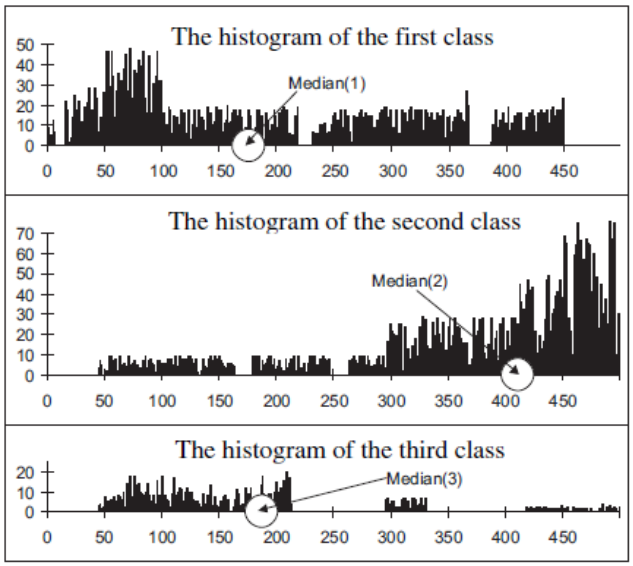
maksimalne razberivosti na osnovu Bulove funkcije razberivosti, tako što je potrebno čekirati ovakvo izračunavanje (*Compute approximate solutions*) i zatim uneti broj aproksimacije (*Hitting fraction*), što se može videti sa slike 18.



Slika 18. Dijalog box za aproksimativno izračunavanje prostih implikanti

U radu sa ovakvom aproksimacijom nije moguće kontrolisati smanjenje konzistentnosti tabele odlučivanja. Druga negativna stvar je što je ovakvom aproksimacijom nekontrolisan kvalitet klasifikacije.

U radu [Nguyen, 2006] predložena je i dokazana mogućnost redukcije određenih tačaka reza za velike baze podataka u odnosu na medijanu (slika 19. preuzeta iz [Nguyen, 2006]).



Slika 19. Histogrami podataka jednog uslovnog atributa za određenu vrednost odluke

Ovakva analiza uključuje posmatranje histograma vrednosti atributa u odnosu na istu vrednost atributa odluke. Zbog toga na slici 19 postoje tri histograma za jedan atribut.

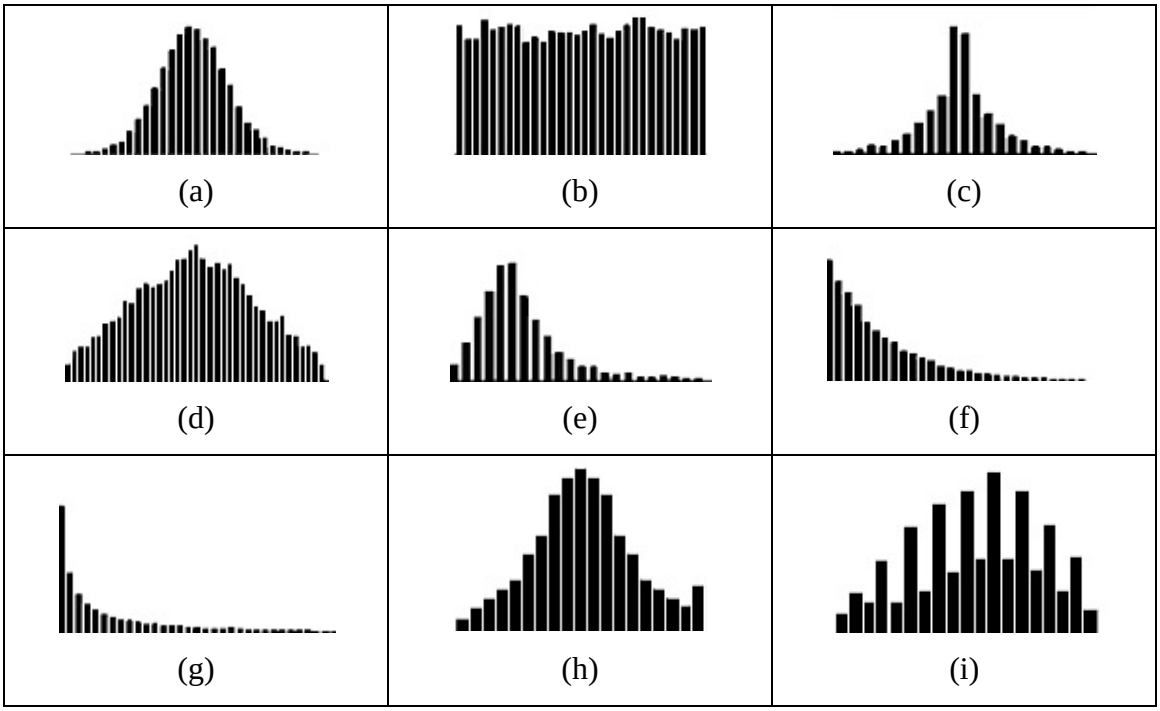
5 ISTRAŽIVANJE

U okviru istraživanja rađena je analiza položaja tačaka reza dobijenih primenom algoritma diskretizacije maksimalne razberivosti i algoritmom baziranim na entropiji nad deset baza koje imaju različite raspodele. Baze nad kojima je urađeno istraživanje preuzete su iz UC Irvine Machine Learning Repository [UCI, 2015]:

1. Iris
2. Blood Transfusion Service Center
3. Banknote Authentication
4. Glass Identification
5. Wilt Data Set
6. Breast Cancer Wisconsin (Original) Data Set
7. Cardiotocography
8. Statlog (Australian Credit Approval)
9. Haberman's Survival Data Set
10. Challenger USA Space Shuttle O-Ring Data Set

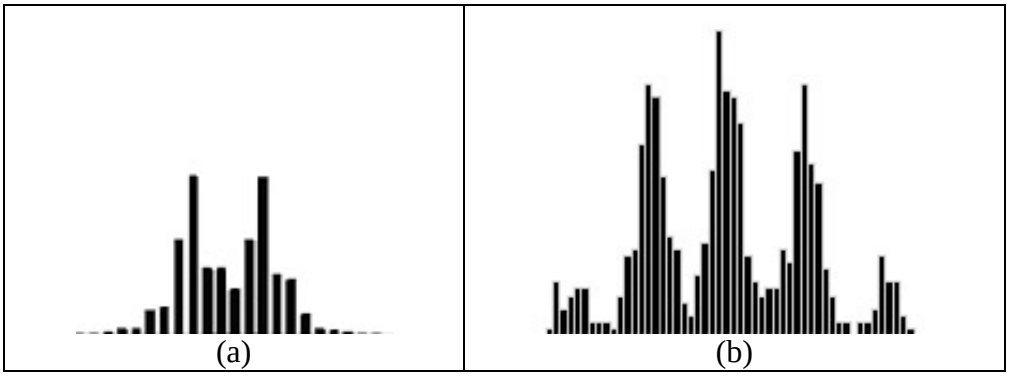
Kao što je navedeno u poglavlju 2, u okviru predmeta istraživanja, sve raspodele su se posmatrale sa stanovišta pripadnosti određenoj grupi raspodela:

1. GRUPA RASPODELA sličnost sa normalnom (slika 20, (a)), uniformnom (slika 20, (b)), leptokurtic (slika 20, (c)), platykurtic (slika 20, (d)), skewed (slika 20, (e)), exponencijalnom (slika 20, (f)), zipf (slika 20, (g)), edge peak (slika 20, (h)), comb (slika 20, (i)) i ostalim raspodelama koje nemaju više izrazitih maksimuma između kojih su periodi minimuma; kod comb raspodele između lokalnih maksimuma postoje minimumi ali bez većih perioda,

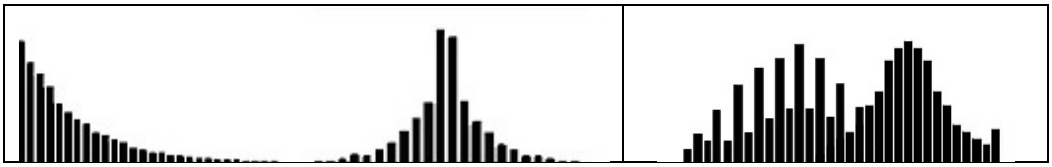


Slika 20. Prva grupa raspodela

2. GRUPA RASPODELA sličnost sa bimodal (slika 21 (a)) i multimodal raspodelom (slika 21 (b)) sa proizvoljnim brojem lokalnih maksimuma; u okviru multimodal raspodele moguća je svaka kombinacija raspodela iz prve grupe sa ili bez perioda prekida kao na slici 22.



Slika 21. Druga grupa raspodela



Slika 22. Primeri netipičnih multimodal raspodela

Na osnovu sličnosti sa nekom od dve grupe raspodela, u tabeli VII navedene su osnovne karakteristike baza i broj atributa koji pripada određenoj grupi raspodela. Za generisanje histograma i proveru pripadnosti određenoj grupi raspodela korišćen je softver EasyFit koji omogućava automatski i ručno generisanje histograma podataka,

kao i analiziranje verovatnoće podataka i određivanje sličnosti sa pogodnim raspodelama [Easy, 2015]. Za svaku bazu uzet je optimalan broj bin-ova histograma [Birge i Rozenholc, 2002].

Baza	Broj uslovnih numeričkih atributa	Broj objekata	Broj uslovnih numeričkih atributa koji imaju raspodelu sličnu 1. grupi	Broj uslovnih numeričkih atributa koji imaju raspodelu sličnu 2. grupi
1. Iris	4	150	1	3
2. Blood Transfusion Service Center	4	748	2	2
3. Banknote Authentication	4	1372	0	4
4. Glass Identification	9	214	3	6
5. Wilt Data Set	5	4339	5	0
6. Breast Cancer Wisconsin (Original) Data Set	9	699	1	8
7. Cardiotocography	22	2126	17	5
8. Statlog (Australian Credit Approval)	6	690	4	2
9. Haberman's Survival Data Set	3	306	2	1
10. Challenger USA Space Shuttle O-Ring Data Set	2	23	0	2

TABELA VII. OSNOVNE KARAKTERISTIKE BAZA NA OSNOVU PRIPADNOSTI ODREĐENOJ GRUPI RASPODELA

Od algoritama diskretizacije koristio se poznat algoritam diskretizacije baziran na entropiji [Fayyad i Irani, 1993] i algoritam maksimalne razberivosti razvijen u okviru teorije grubih skupova [Nguyen, 2006].

Za diskretizaciju podataka koristio se sistem Rosetta koji je napravljen nakon razvitka algoritma maksimalne razberivosti [Øhrn, 1998] i on omogućuje rad sa izabranim diskretizacijama. Algoritam maksimalne razberivosti, u literaturi često označen po teoriji na kojoj je baziran, Boolean reasoning algorithm, u sistemu Rosetta ima naziv *BROrthogonalScaler*. To je pohlepni (*greedy*) algoritam za određivanje minimalnog skupa pokrivanja (*minimal set covering*) objekata iz različitih klasa atributa odluke. On je direktna implementacija algoritma koji su opisali Nguyen i Skowron [Nguyen, 1996], baziran na kombinovanju početnog ukupnog skupa kandidata tačaka reza Boolean reasoning procedurom koja odbacuje određene tačke reza do malog podskupa. Dobijen mali podskup tačaka reza je minimalan skup tačaka reza koji čuva razberivost svojstvenu sistemima odlučivanja. Treba naglasiti da u slučaju da ako se neki atribut ne diskretizuje (nema ni jednu tačku reza), onda se uzimaju sve njegove vrednosti, odnosno podintervalima (dobijenim tačkama reza za diskretizovane atribute) se

pridružuju sve različite vrednosti nediskretizovanih atributa. Sa stanovišta analize tačaka reza to znači da se tačkama reza diskretizovanih atributa pridružuju sve tačke kandidati reza za nediskrezovane atribute.

Algoritam baziran na entropiji u sistemu Rosetta ima naziv *EntropyScaler*. Implementacija ovog algoritma je opisana u [Dougherty i dr., 1995], a bazirana je na rekurzivnoj particiji skupa vrednosti svakog atributa tako da je lokalna mera entropije optimalna. Princip dužina minimalnih karakteristika definiše kriterijum koji zaustavlja proces particije.

5.1 UTICAJ RASPODELE PODATAKA NA ALGORITME DISKRETIZACIJE

Nad bazama je urađeno ispitivanje uticaja raspodele podataka na dva algoritma diskretizacije i to na algoritam baziran na entropiji i algoritam maksimalne razberivosti. Sve slike histograma sa označenim tačkama reza nalaze se u Dodacima. Praćena je veličina dobijenih podintervala u odnosu na ukupni interval kao i broj objekata koji pripada određenom podintervalu. Naglasak je dat na uočavanju položaja tačaka reza u odnosu na jednu od dve grupe raspodela.

1. Analiza baze Iris - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Tačke reza dobijene algoritmom maksimalne razberivosti dele čitav interval uglavnom na veće podintervale bez obzira na tip raspodele. Primećuje se da se tačke reza kod multimodal raspodela nalaze nakon dela koji odgovara raspodeli iz 1. grupe.
- Tačke reza dobijene algoritmom baziranim na entropiji dele čitav interval na veći levi interval i na nekoliko manjih podintervala, naročito u situacijama kada raspodela nije slična sa normalnom raspodelom.

2. Analiza baze Blood Transfusion Service Center - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Raspodele prvog i četvrtog atributa baze Blood Transfusion Service Center su vrlo nepravilne, sa puno izrazitih lokalnih maksimuma. Na osnovu algoritma diskretizacije maksimalne razberivosti, dobijene tačke reza kod prvog i četvrtog atributa dele čitav interval na podintervale oko lokalnih maksimuma. Drugi atribut iako ima pravilnu raspodelu je podeljen na podintervale u skladu sa brojnošću tačaka reza prvog i četvrtog atributa.
- Na osnovu algoritma diskretizacije baziranog na entropiji, dobijene tačke reza dele čitav interval na veći levi interval i na nekoliko manjih podintervala, naročito u situacijama kada raspodela nije slična sa normalnom raspodelom, ekponencijalnom raspodelom ili njima sličnim raspodelama.

3. Analiza baze Banknote Authentication - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- S' obzirom na broj instanci ove baze, očekuje se vrlo nepravilna raspodela. U slučaju diskretizacije algoritmom maksimalne razberivosti, može se primetiti da se tačke reza kod multimodal raspodele uglavnom nalaze posle lokalnih maksimuma koji su značajno veći od ostalih lokalnih maksimuma ili u slučajevima kada nakon lokalnih maksimuma postoji lokalni minimum – to su tačke na histogramu koje se nalaze nakon dela histograma koji je sličan raspodelama iz 1. grupe. Pri krajevima intervala raspodele nema tačaka reza. Podintervali uglavnom sadrže podjednak broj objekata.
- Za bazu sa preko 1000 instanci i sa vrlo napravilnom raspodelom, rezultat diskretizacije algoritmom baziranim na entropiji daje mnogo više tačaka reza u odnosu na algoritam maksimalne razberivosti. Dobijene tačke reza u većini slučajeva dele interval tako da deo histograma sličan sa normalnom raspodelom

ima veći broj tačaka reza. Primećuje se disproporcija u veličini podintervala, a naročito postoji disproporcija u broju objekata po pojedinim podintervalima.

4. Analiza baze Glass Identification - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Skoro kod svih atributa, tačke reza (dobijene algoritmom diskretizacije maksimalne razberivosti) dele histograme multimodal raspodela tako da su delovi histograma dobijeni deobom na osnovu tačaka reza slični raspodelama iz 1. grupe. Sa malim izuzetkom, na krajevima intervala raspodele nema tačaka reza.
- Rezultat diskretizacije algoritmom baziranim na entropiji daje mnogo više tačaka reza u odnosu na algoritam maksimalne razberivosti. Primećuje se disproporcija u veličini podintervala, a naročito postoji disproporcija u broju objekata po pojedinim podintervalima.

5. Analiza baze Wilt – uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Svi histogrami pripadaju 1. grupi i predstavljaju varijantu normalne raspodele. Za veoma veliki broj objekata, pravilnost raspodele je rezultovala generisanjem malog broja tačaka reza koje uglavnom dele histogram na podjednake podintervale.
- Postoji mnogostuko veći broj tačaka reza nego kod algoritma maksimalne razberivosti. Primećuje se disproporcija u veličini podintervala, a naročito postoji disproporcija u broju objekata po pojedinim podintervalima.

6. Analiza baze Breast Cancer Wisconsin (Original) Data Set - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Tačke reza (dobijene algoritmom diskretizacije maksimalne razberivosti) dele histograme multimodal raspodela tako da su delovi histograma dobijeni deobom na osnovu tačaka reza slični raspodelama iz 1. grupe. Na krajevima intervala histograma nema tačaka reza.
- Rezultat diskretizacije algoritmom baziranim na entropiji daje podjednak broj tačaka reza u odnosu na algoritam maksimalne razberivosti – ovo je ređi slučaj baze sa malim brojem različitih vrednosti na histogramu, kada se primećuje da je broj tačaka reza podjednak. Tačke reza se uglavnom nalaze na krajevima intervala histograma.

7. Analiza baze Cardiotocography – uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- S' obzirom na veličinu baze postoji prilično mali broj tačaka reza dobijenih algoritmom diskretizacije maksimalne razberivosti. Kod zipf raspodela uočljivo je nepostojanje tačaka reza, što je intuitivno u skladu sa funkcijom razberivosti.
- U odnosu na tačke reza dobijene algoritmom maksimalne razberivosti, postoji značajno više tačaka reza dobijene algoritmom baziranim na entropiji. Tačke reza se kod obe rupe raspodela uglavnom nalaze na krajevima intervala i čitav

interval dele na nekoliko velikih podintervala i mnogo manje zanemarljivo malih podintervala na krajevima intervala. Kod raspodela koje u velikom procentu ima samo jednu vrednost, takođe postoji nakoliko tačaka reza koje generišu podintervale sa zanemarljivo malim brojem objekata.

8. Analiza baze Statlog (Australian Credit Approval) - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Tačke reza (dobijene algoritmom diskretizacije maksimalne razberivosti) dele histograme multimodal raspodela tako da su delovi histograma dobijeni deobom na osnovu tačaka reza slični raspodelama iz 1. grupe. Zbog toga na histogramima sa većim brojem „unimodal“ raspodela u okviru multimodal raspodela postoji veći broj tačaka reza nego kod histograma sličnim raspodelama iz 1. grupe. Nema tačaka reza koje određuju mali broj objekata.
- Rezultat diskretizacije algoritmom baziranim na entropiji daje mnogo veći broj tačaka reza u odnosu na algoritam maksimalne razberivosti kod svih multimodal raspodela. Dobijeni podintervali su sa vrlo neproporcionalnim brojem objekata. Postoji velik broj tačaka reza koje se nalaze na krajevima intervala histograma.

9. Analiza baze Haberman's Survival Data Set - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Drugi atrubut ima uniformnu raspodelu, pa se tačke reza dobijene algoritmom maksimalne razberivosti nalaze nakon skoro svake vrednosti na histogramu. To je razlog što samo neke od tačaka reza ostalih atributa dele histograme tako da su delovi histograma dobijeni deobom na osnovu tih tačaka reza slični raspodelama iz 1. grupe.
- Rezultat diskretizacije algoritmom baziranim na entropiji daje tačke reza koje dele histogram na podintervale sa jednim većim i nekoliko manjih podintervala sa neproporcinalnim brojem objekata. Tačke reza se uglavnom nalaze na krajevima intervala histograma.

10. Analiza baze Challenger USA Space Shuttle O-Ring Data Set - uticaj raspodele podataka na algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji:

- Za mali broj objekata tačke reza dobijene algoritmom maksimalne razberivosti dele histogram na više podintervala precizno u skladu sa funkcijom razberivosti. Histogram multimodal raspodele je tačakama reza podeljen na raspodelu sličnu 1. grupi.
- Rezultat diskretizacije algoritmom baziranim na entropiji je samo jedna tačka reza jednog atributa.

Zaključak poglavlja 5.1

Na osnovu analiziranih histograma atributa baza i položaja tačaka reza dobijenih diskretizacijom algoritmom maksimalne razberivosti i algoritmom diskretizacije baziranog na entropiji, može se zaključiti sledeće:

- tačke reza dobijene algoritmom maksimalne razberivosti se nalaze nakon lokalnih maksimuma ili nakon dela histograma multimodal raspodele koji je sličan nekoj raspodeli iz 1. grupe;
- tačke reza dobijene algoritmom diskretizacije baziranom na entropiji dele interval histograma vrlo nepravilno naročito u slučaju multimodal raspodela (postoji velika disproporcija u broju objekata koje određuju tačke reza).

5.2 ANALIZA KLASIFIKACIJA NA OSNOVU RASPODELA PODATAKA

U ovom poglavlju urađena je analiza rezultata klasifikacija u odnosu na već analizirane raspodele iz prethodnog poglavlja i dela Dodaci. Nakon svake tabele sa rezultatima jedne baze opisane su veze sa karakterističnim raspodelama određenog/određenih atributa u cilju pronalaženja veza između uticaja određene tačke reza na rezultat klasifikacije.

Za deset baza iz poglavlja 5.1 i dobijenih tačaka reza na osnovu primenjenih diskretizacija (algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji), urađena je klasifikacija u sistemu Rosetta. Postupak klasifikacije urađen je na sledeći način:

1. izvršena je podela baze na dva dela tako da su u jednu polovinu baze smešteni objekti sa neparnim rednim brojem, a u drugu polovinu baze objekti sa parnim rednim brojem,
2. izračunat je redukt prve polovine baze na osnovu Džonsonovog algoritma,
3. generisana su IF THEN pravila prve polovine baze (na osnovu prethodno dobijenog redukta),
4. urađena je klasifikacija druge polovine baze na osnovu IF THEN pravila prve polovine baze.

Rezultat klasifikacije je dat matricom konfuzije koja opisuje broj objekata koji su klasifikovani na osnovu pojedinih klasa odluke, kao i broj neklasifikovanih objekata. Time je moguće pratiti broj objekata koji su se tačno i netačno klasifikovali.

1. Analiza klasifikacije baze Iris

Rezultati klasifikacija dati su matricama konfuzije na slikama 23. i 24., a analiza klasifikacije tabelom VIII.

		Predicted				
		1	2	3	Undefined	
Actual	1	26	0	0	0	1.0
	2	0	26	2	0	0.928571
	3	0	0	20	1	0.952381
	Undefined	0	0	0	0	Undefined
		1.0	1.0	0.909091	0.0	0.96

Slika 23. Matrica konfuzije baze Iris diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted				
Actual		1	2	3	Undefined	
	1	26	0	0	0	1.0
	2	0	12	0	16	0.428571
	3	0	2	12	7	0.571429
	Undefined	0	0	0	0	Undefined
		1.0	0.857143	1.0	0.0	0.666667

Slika 24. Matrica konfuzije baze Iris diskretizovane na osnovu algoritma baziranog na entropiji

1. Iris	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a3, a4}	{a1, a3}
Broj pravila koja su generisana na osnovu redukta	6	26
Broj nepreciznih pravila	0	0
Procenat nepreciznih pravila	0%	0%
Ukupan rezultat matrice konfuzije	0.96	0.67

TABELA VIII. IRIS - ANALIZA KLASIFIKACIJE

Iris, MD: Ukupan rezultat klasifikacije iznosi 0.96 što predstavlja odličan rezultat. Za bazu koja ima relativno mali broj tačaka reza dobijenih algoritmom maksimalne razberivosti (Dodaci 1), i za očekivati je da ima dobar rezultat klasifikacije, jer su objekti podeljeni u velike podskupove na osnovu relacije razberivosti.

Iris, E: Ukupan rezultat klasifikacije iznosi 0.67 što je značajno lošiji rezultat u odnosu na MD. Tačke reza dobijene algoritmom baziranim na entropiji iako uglavnom daju prvi veći interval (Dodaci 1), zbog većeg broja preostalih značajno manjih intervala (intervala sa disproporcionalno manjim brojem objekata), imaju veći broj IF THEN pravila, a time i lošiji rezultat klasifikacije.

2. Analiza klasifikacije baze Blood Transfusion Service Center

Rezultati klasifikacija dati su matricama konfuzije na slikama 25. i 26., a analiza klasifikacije tabelom IX.

		Predicted			
		0	1	Undefined	
Actual	0	119	3	164	0.416084
	1	25	9	54	0.102273
	Undefined	0	0	0	Undefined
		0.826389	0.75	0.0	0.342246

Slika 25. Matrica konfuzije baze Blood Transfusion Service Center diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		0	1	Undefined	
Actual	0	275	3	8	0.961538
	1	84	2	2	0.022727
	Undefined	0	0	0	Undefined
		0.766017	0.4	0.0	0.740642

Slika 26. Matrica konfuzije baze Blood Transfusion Service Center diskretizovane na osnovu algoritma baziranog na entropiji

2. Blood Transfusion Service Center	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a1, a2, a4}	{a1, a2, a4}
Broj pravila koja su generisana na osnovu redukta	266	32
Broj nepreciznih pravila	11	10
Procenat nepreciznih pravila	4%	31%
Ukupan rezultat matrice konfuzije	0.34	0.74

TABELA IX. BLOOD TRANSFUSION SERVICE CENTER - ANALIZA KLASIFIKACIJE

Blood Transfusion Service Center, MD: Ukupan rezultat klasifikacije iznosi 0.34 što je loš rezultat. Za bazu koja ima 748 objekata i veći broj tačaka reza dobijenim algoritmom maksimalne razberivosti (Dodaci 2), značajno je primetiti uticaj vrlo nepravilne raspodele kod četvrtog atributa, sa puno izraženih podintervala koji su svaki pojedinačno u određenoj meri slični normalnoj raspodeli. Takođe, na histogramu četvrtog atributa se može primetiti da ne postoje veća rastojanja oko tačaka reza, što znači da je moguće da se podaci ne svrstavaju po jasnoj pripadnosti određnom podintervalu.

Blood Transfusion Service Center, E: Ukupan rezultat klasifikacije iznosi 0.74 što je duplo bolje u odnosu na MD. Bitno je uočiti da četvrti atribut sa vrlo nepravilnom

raspodelom ima samo četiri tačke reza (Dodaci 2), i da je zbog toga dobijen značajno manji broj IF THEN pravila u odnosu na podatke diskretizovane algoritmom maksimalne razberivosti (32 pravila prema 265 pravila). To je rezultovalo većim procentom nepreciznih pravila na osnovu kojih je dobijen bolji ukupni rezultat matrice konfuzije. Moglo bi se zaključiti da se za oko 6 puta nepreciznu klasifikaciju dobija duplo bolji rezultat klasifikacije.

3. Analiza klasifikacije baze Banknote Authentication

Rezultati klasifikacija dati su matricama konfuzije na slikama 27. i 28., a analiza klasifikacije tabelom X.

		Predicted			
		0	1	Undefined	
Actual	0	362	0	14	0.962766
	1	2	301	7	0.970968
	Undefined	0	0	0	Undefined
		0.994505	1.0	0.0	0.966472

Slika 27. Matrica konfuzije baze Banknote Authentication diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		0	1	Undefined	
Actual	0	181	0	195	0.481383
	1	0	131	179	0.422581
	Undefined	0	0	0	Undefined
		1.0	1.0	0.0	0.454811

Slika 28. Matrica konfuzije baze Banknote Authentication diskretizovane na osnovu algoritma baziranog na entropiji

3. Banknote Authentication	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a1, a2, a3}	{a1, a2, a3}
Broj pravila koja su generisana na osnovu redukta	67	432
Broj nepreciznih pravila	0	0
Procenat nepreciznih pravila	0%	0%
Ukupan rezultat matrice konfuzije	0.97	

TABELA X. BANKNOTE AUTHENTICATION - ANALIZA KLASIFIKACIJE

Banknote Authentication, MD: Ukupan rezultat klasifikacije iznosi 0.97 što je odličan rezultat. Za bazu koja ima 1372 objekata i nepravilne raspodele (svi atributi imaju

multimodal raspodele, Dodaci 3), na osnovu nekoliko tačaka reza, samo dva objekta su nepravilno klasifikovana.

Banknote Authentication, E: Ukupan rezultat klasifikacije iznosi 0.45 što je značajno lošije u odnosu na MD. Zbog vrlo nepravilnih multimodal raspodela (Dodaci 3) broj tačaka reza je nekoliko stotina puta veći nego kod MD, pa je rezultat je loš zbog velikog broja pravila koja se odnose na male intervale.

4. Analiza klasifikacije baze Glass Identification

Rezultati klasifikacija dati su matricama konfuzije na slikama 29. i 30., a analiza klasifikacije tabelom XI.

		Predicted							
		1	2	3	5	6	7	Undefined	
Actual	1	15	0	0	0	0	0	23	0.394737
	2	0	13	0	0	0	0	25	0.342105
	3	0	0	1	0	0	0	5	0.166667
	5	0	0	0	0	0	0	5	0.0
	6	0	0	0	0	0	0	3	0.0
	7	0	0	0	0	0	7	10	0.411765
	Undefined	0	0	0	0	0	0	0	Undefined
		1.0	1.0	1.0	Undefined	Undefined	1.0	0.0	0.336449

Slika 29. Matrica konfuzije baze Glass Identification diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted							
		1	2	3	5	6	7	Undefined	
Actual	1	17	2	0	0	0	0	19	0.447368
	2	1	9	0	0	0	0	28	0.236842
	3	1	1	0	0	0	0	4	0.0
	5	0	0	0	0	0	0	5	0.0
	6	0	0	0	0	0	0	3	0.0
	7	0	0	0	0	0	0	17	0.0
	Undefined	0	0	0	0	0	0	0	Undefined
		0.894737	0.75	Undefined	Undefined	Undefined	Undefined	0.0	0.242991

Slika 30. Matrica konfuzije baze Glass Identification diskretizovane na osnovu algoritma baziranog na entropiji

4. Glass Identification	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a2, a3, a4, a5, a6, a7, a8, a10}	{a2, a3, a5, a7, a8}
Broj pravila koja su generisana na osnovu redukta	88	96
Broj nepreciznih pravila	0	2
Procenat nepreciznih pravila	0%	2%
Ukupan rezultat matrice konfuzije	0.34	0.24

TABELA XI. GLASS IDENTIFICATION - ANALIZA KLASIFIKACIJA

Glass Identification, MD: Ukupan rezultat klasifikacije iznosi 0.34 što je loš rezultat. Za bazu koja ima 214 objekata i devet atributa od kojih atributi a2, a3, a4, a5, a6, a7 i a8 imaju više lokalnih maksimuma, značajno je primetiti relativno mali broj tačaka reza sa jedne strane i velik broj atributa koji čine redukt sa druge strane. Time se generišu pravila koja imaju veći broj uslova, što dovodi do toga da se ne može klasifikovati značajan broj objekata iz test skupa (Dodaci 4).

Glass Identification, E: Ukupan rezultat klasifikacije iznosi 0.24 i predstavlja još lošiji rezultat u odnosu na MD. Može se uočiti da je veliki broj tačaka reza (Dodaci 4) generisao još veći broj pravila. Samo dva pravila su neprecizna i predstavljaju zanemarljiv procenat u odnosu na ukupan broj pravila, tako da ne mogu bitnije da utiču na rezultat.

U slučaju baze Glass Identification može se primetiti da kod velikog broja atributa sa nepravilnim raspodelama, gde se redukt sastoji od većeg broja atributa, kod obe diskretizacije postoji velik broj tačnih pravila sa puno uslova (IF deo pravila) na osnovu kojih se ne može dobiti veliki procenat uspešne klasifikacije.

5. Analiza klasifikacije baze Wilt Data Set

Rezultati klasifikacija dati su matricama konfuzije na slikama 31. i 32., a analiza klasifikacije tabelom XII.

		Predicted			
		w	n	Undefined	
Actual	w	15	0	18	0.454545
	n	4	2071	61	0.969569
	Undefined	0	0	0	Undefined
		0.789474	1.0	0.0	0.961734

Slika 31. Matrica konfuzije baze Wilt Data Set diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		w	n	Undefined	
Actual	w	1	1	31	0.030303
	n	0	2134	2	0.999064
	Undefined	0	0	0	Undefined
		1.0	0.999532	0.0	0.984325

Slika 32. Matrica konfuzije baze Wilt Data Set diskretizovane na osnovu algoritma baziranog na entropiji

5. Wilt Data Set	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{GLCM_pan,Mean_Green, Mean_Red, Mean_NIR}	{SD_pan}
Broj pravila koja su generisana na osnovu redukta	215	113
Broj nepreciznih pravila	0	0
Procenat nepreciznih pravila	0%	0%
Ukupan rezultat matrice konfuzije	0.96	0.98

TABELA XII. WILT DATA SET - ANALIZA KLASIFIKACIJA

Wilt Data Set, MD: Od ukupno pet uslovnih atributa u reduktu se nalaze čak četiri. Za relativno velik broj elemenata redukta rezultat klasifikacije je odličan: 0.96. Sve raspodele atributa pripadaju 1. grupi i sve su veoma pravilne normalne raspodele (Dodaci 5). Takve veoma pravilne raspodele rezultovale su pravilnošću u odnosu na razberivost.

Wilt Data Set, E: Ukupan rezultat klasifikacije iznosi 0.98 što je vrlo malo bolje u odnosu na MD. Bitno je uočiti da je klasifikacija rađena na osnovu pravila generisanih nad reduktom sa samo jednim atributom (Dodaci 5)..

6. Analiza klasifikacije baze Breast Cancer Wisconsin (Original) Data Set

Rezultati klasifikacija dati su matricama konfuzije na slikama 33. i 34., a analiza klasifikacije tabelom XIII.

		Predicted			
		2	4	Undefined	
Actual	2	221	2	10	0.948498
	4	2	47	67	0.405172
	Undefined	0	0	0	Undefined
		0.991031	0.959184	0.0	0.767908

Slika 33. Matrica konfuzije baze Breast Cancer Wisconsin diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		2	4	Undefined	
Actual	2	224	6	3	0.961373
	4	10	22	84	0.189655
	Undefined	0	0	0	Undefined
		0.957265	0.785714	0.0	0.704871

Slika 34. Matrica konfuzije baze Breast Cancer Wisconsin diskretizovane na osnovu algoritma baziranog na entropiji

6. Breast Cancer Wisconsin (Original) Data Set	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a2, a4, a5, a6, a9, a10}	{a2, a3, a4, a5, a6, a7, a9, a10}
Broj pravila koja su generisana na osnovu redukta	129	103
Broj nepreciznih pravila	0	5
Procenat nepreciznih pravila	0%	5%
Ukupan rezultat matrice konfuzije	0.77	0.70

TABELA XIII. BREAST CANCER WISCONSIN (ORIGINAL) DATA SET - ANALIZA KLASIFIKACIJA

Breast Cancer Wisconsin (Original) Data Set, MD: Ukupan rezultat klasifikacije iznosi 0.77 što je dobar rezultat. Za bazu koja ima 699 objekata i relativno mali broj različitih vrednosti atributa, postoji mali broj tačaka reza: dva tributa imaju po tri tačke reza, dok ostali imaju manje (Dodaci 6). Za 2. atribut Clump Thickness, tri tačke reza dele čitav interval na ujednačene podintervale. Ukupno posmatrajući histograme, bez obzira na velik broj objekata, zbog postojanja malog broja vrednosti za svaki atribut, ovo je primer baze sa težim određivanjem sličnosti određenoj raspodeli.

Breast Cancer Wisconsin (Original) Data Set, E: Ukupan rezultat klasifikacije iznosi 0.70 što je malo lošije u odnosu na prethodno analiziranu matricu konfuzije (slika 33). Za 2. atributa Clump Thickness postoji samo jedna tačka reza koja deli iterval na dva disproporcionalna podintervala što je drastično u odnosu na prethodni slučaj.

7. Analiza klasifikacije baze Cardiotocography

Rezultati klasifikacija dati su matricama konfuzije na slikama 35. i 36., a analiza klasifikacije tabelom XIV.

		Predicted				
		1	2	3	Undefined	
Actual	1	324	4	0	501	0.390832
	2	4	62	0	79	0.427586
	3	0	1	30	58	0.337079
	Undefined	0	0	0	0	Undefined
		0.987805	0.925373	1.0	0.0	0.391345

Slika 35. Matrica konfuzije baze Cardiotocography diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted				
		1	2	3	Undefined	
Actual	1	279	5	0	545	0.33655
	2	6	31	0	108	0.213793
	3	0	0	45	44	0.505618
	Undefined	0	0	0	0	Undefined
		0.978947	0.861111	1.0	0.0	0.333961

Slika 36. Matrica konfuzije baze Cardiotocography diskretizovane na osnovu algoritma baziranog na entropiji

7. Cardiotocography	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{MSTV, Max, CLASS}	{AC, UC, MLTV, Width, Min, Max, Nmax, Variance, Tendency, CLASS}
Broj pravila koja su generisana na osnovu redukta	811	811
Broj nepreciznih pravila	0	1
Procenat nepreciznih pravila	0%	0%
Ukupan rezultat matrice konfuzije	0.39	0.33

TABELA XIV. CARDIOTOCOGRAPHY - ANALIZA KLASIFIKACIJA

Cardiotocography, MD: Ukupan rezultat klasifikacije iznosi 0.39. Za bazu koja ima 2126 objekata, 22 uslovna atributa i relativno mali broj tačaka reza, može se primetiti da najveći uticaj čine tačke reza 22. atributa CLASS, koje dele interval na pet podintervala – ovaj atribut ima najnepravilniju multimodal raspodelu. Time se potvrđuje da je mišljenje tri eksperta koji su napisali atribut odluke NSP u skladu sa atributom CLASS (Dodaci 7). Rezultat je osrednji zbog vrlo razliitih raspodela.

Cardiotocography, E: Ukupan rezultat klasifikacije iznosi 0.33 što je lošije u odnosu na algoritam MD. Iako tačke reza na osnovu diskretizacije entropijom dele histograme

vrlo nepravilno, rezultat je osrednji zbog postojanja većih podintervala kod nekih histograma (bez postojanja većih podintervala rezultat bi bio lošiji).

8. Analiza klasifikacije baze Statlog (Australian Credit Approval)

Rezultati klasifikacija dati su matricama konfuzije na slikama 37. i 38., a analiza klasifikacije tabelom XIV.

		Predicted			
		0	1	Undefined	
Actual	0	23	4	169	0.117347
	1	3	3	143	0.020134
	Undefined	0	0	0	Undefined
		0.884615	0.428571	0.0	0.075362

Slika 37. Matrica konfuzije baze Statlog (Australian Credit Approval) diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		0	1	Undefined	
Actual	0	107	16	73	0.545918
	1	17	28	104	0.187919
	Undefined	0	0	0	Undefined
		0.862903	0.636364	0.0	0.391304

Slika 38. Matrica konfuzije baze Statlog (Australian Credit Approval) diskretizovane na osnovu algoritma baziranog na entropiji

8. Statlog (Australian Credit Approval)	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a2, a3, a5, a6, a7, a10, a11, a13, a14}	{a1, a2, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14}
Broj pravila koja su generisana na osnovu redukta	335	245
Broj nepreciznih pravila	0	11
Procenat nepreciznih pravila	0%	4%
Ukupan rezultat matrice konfuzije	0.08	0.39

TABELA XV. STATLOG (AUSTRALIAN CREDIT APPROVAL) - ANALIZA KLASIFIKACIJA

Statlog (Australian Credit Approval), MD: Od ukupno 14 uslovnih atributa u reduktu se nalazi 9. Velik broj elemenata redukta i raspodele iz 2. grupe dali su loš rezultat klasifikacije (Dodaci 8).

Statlog (Australian Credit Approval), E: Za još veći broj elemenata redukta nego kod MD algoritma, ukupan rezultat klasifikacije je značajno bolji u odnosu na MD (Dodaci 8). Za to su najviše zaslužni veliki intervali koje su generisale tačke reza.

9. Analiza klasifikacije baze Haberman's Survival Data Set

Rezultati klasifikacija dati su matricama konfuzije na slikama 39. i 40., a analiza klasifikacije tabelom XVI.

		Predicted			
		1	2	Undefined	
Actual	1	20	2	90	0.178571
	2	3	3	35	0.073171
	Undefined	0	0	0	Undefined
		0.869565	0.6	0.0	0.150327

Slika 39. Matrica konfuzije baze Haberman's Survival Data Set diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		1	2	Undefined	
Actual	1	112	0	0	1.0
	2	37	3	1	0.073171
	Undefined	0	0	0	Undefined
		0.751678	1.0	0.0	0.751634

Slika 40. Matrica konfuzije baze Haberman's Survival Data Set diskretizovane na osnovu algoritma baziranog na entropiji

8. Haberman's Survival Data Set	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a1, a2, a3}	{a1, a3}
Broj pravila koja su generisana na osnovu redukta	141	9
Broj nepreciznih pravila	0	3
Procenat nepreciznih pravila	0%	33%
Ukupan rezultat matrice konfuzije	0.15	0.75

TABELA XVI. HABERMAN'S SURVIVAL DATA SET - ANALIZA KLASIFIKACIJA

Haberman's Survival Data Set, MD: U reduktu skupa se nalaze sva tri uslovna atributa. Prvi atribut ima vrlo nepravilnu multimodal raspodelu i relativno mnogo tačaka reza, što je uticalo da i ostala dva atributa imaju relativno mnogo tačaka reza (Dodaci 9). To je razlog lošeg rezultata klasifikacije.

Haberman's Survival Data Set, E: Dva atributa su u reduktu i velik je procenat nepreciznih pravila. Rezultat klasifikacije je mnogo bolji nego kod MD algoritma. Za to su najviše zaslužni veliki intervali koje su generisale tačke reza i neprecizna pravila (Dodaci 9).

10. Analiza klasifikacije baze Challenger USA Space Shuttle O-Ring Data Set

Rezultati klasifikacija dati su matricama konfuzije na slikama 41. i 42., a analiza klasifikacije tabelom XVII.

		Predicted			
		0	1	Undefined	
Actual	0	7	0	2	0.777778
	1	0	1	2	0.333333
	Undefined	0	0	0	Undefined
		1.0	1.0	0.0	0.666667

Slika 41. Matrica konfuzije baze Challenger USA Space Shuttle O-Ring Data Set diskretizovane na osnovu algoritma maksimalne razberivosti

		Predicted			
		0	1	Undefined	
Actual	0	5	2	2	0.555556
	1	1	2	0	0.666667
	Undefined	0	0	0	Undefined
		0.833333	0.5	0.0	0.583333

Slika 42. Matrica konfuzije baze Challenger USA Space Shuttle O-Ring Data Set diskretizovane na osnovu algoritma baziranog na entropiji

10. Challenger USA Space Shuttle O-Ring Data Set	MD: Rezultati na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti	E: Rezultati na osnovu podataka diskretizovanih algoritmom baziranim na entropiji
Redukt prve polovine diskretizovane baze	{a3}	{a3, a4}
Broj pravila koja su generisana na osnovu redukta	4	4
Broj nepreciznih pravila	0	1
Procenat nepreciznih pravila	0%	25%
Ukupan rezultat matrice konfuzije	0.67	0.58

TABELA XVII. CHALLENGER USA SPACE SHUTTLE O-RING DATA SET - ANALIZA KLASIFIKACIJA

Challenger USA Space Shuttle O-Ring Data Set, MD: Za mali broj objekata evidentan je rizik od eksplozije. Bez obzira na raspodelu (Dodaci 10), na osnovu malog redukta dobila su se jasna pravila.

Challenger USA Space Shuttle O-Ring Data Set, E: U reduktu su dva atributa (temperatura i pritisak), ali je velik procenat neprecinih pravila. Uprkos tome rezultat klasifikacije je loš.

Zaključak poglavlja 5.2

Na osnovu analiziranih redukata baza, broja dobijenih pravila odlučivanja, procenta nepreciznih pravila, kao i rezultata matrice konfuzije za oba posmatrana algoritma, može se zaključiti sledeće:

- nad podacima sa multimodal raspodelama koji su diskretizovani algoritmom maksimalne razberivosti primećuje se veći redukt (sa više atributa) što rezultuje većim brojem pravila i lošijom klasifikacijom u odnosu na podatke sa raspodelama iz 1. grupe;
- nad podacima sa multimodal raspodelama koji su diskretizovani algoritmom baziranim na entropiji primećuje se veći procenat neprecinih pravila (u odnosu na algoritam maksimalne razberivosti) koja čine da je rezultat klasifikacije bolji.

5.3 REDUKCIJA TAČAKA REZA DOBIJENIH ALGORITMOM MAKSIMALNE RAZBERIVOSTI

Za razradu ideje redukcije tačaka reza uzeće se dve baze i to baza Iris koja ima odličan rezultat klasifikacije i baze Blood Transfusion Service Center koja ima loš rezultat klasifikacije. I jedna i druga baza imaju raspodele iz obe grupe raspodela.

5.3.1 Redukcija tačaka reza dobijenih algoritmom maksimalne razberivosti nad bazom koja ima dobar rezultat klasifikacije

Ako se posmatraju tačke reza baze Iris (Dodaci 1) može se uočiti mali podinterval 2. atributa sepal width koji je dobijen na početku intervala: [*, 2.3) koji deluje najmanje značajno jer ima najmanji broj objekata. Ako se ova tačka reza izbaci, redukt se ne menja a ne menja se ni rezultat klasifikacije, odnosno matrica konfuzije ostaje ista. Ako se izbace obe tačke reza drugog atributa ni tada se ne menja redukt a ni rezultat klasifikacije, odnosno matrica konfuzije ponovo ostaje ista. Isto će se desiti ako izbacimo sve tačke reza prvog i drugog atributa. Pošto su pravila klasifikacije bazirana na reduktu skupa, a redukt baze Iris je {a3, a4} (poglavlje 5.2), to je razlog zašto izbacivanje bilo koje tačke atributa koji se na sadrži u reduktu, ne utiče na redukt skupa. Na početku istraživanja naslućivalo se da izbacivanje određenih tačaka reza ne utiče na redukt skupa, ali je pretpostavka postavljena na osnovu empirijskih posmatranja.

Ako se donese odluka o redukciji tačke reza atributa koji se nalazi u reduktu, tada se može očekivati promena redukta i/ili ostalih rezultata klasifikacije. Uz redukciju tačke reza atributa koji pripada reduktu, može se uraditi redukcija određenih tačaka reza atributa koji ne pripadaju reduktu. Ideja je da se prati raspodela podataka, i da se redukcija tačke radi na osnovu veličine podintervala koji određuje i položaja na histogramu.

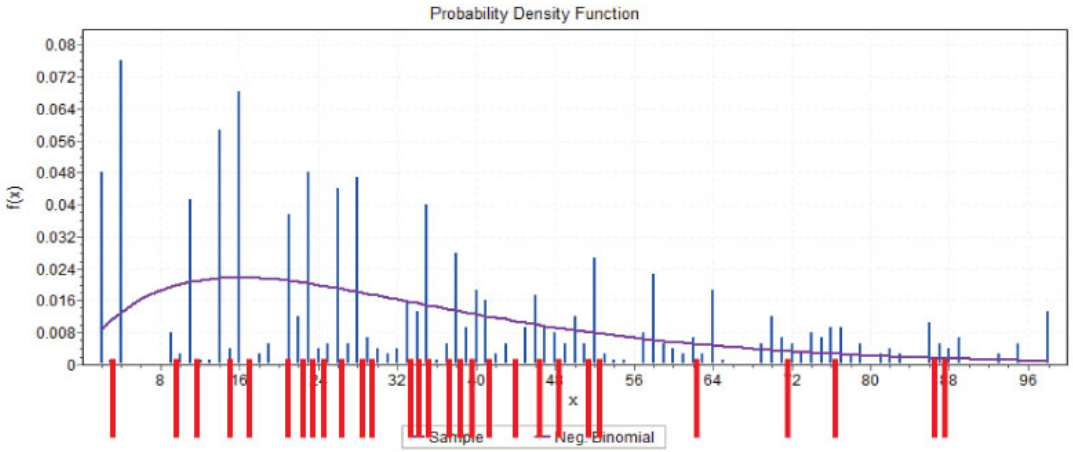
Ponovo će se izbaciti tačka reza 2. atributa sepal width, tačka 22.5 jer generiše podinterval sa najmanjim brojem objekata. Pored nje će se izbaciti i tačka reza 4. atributa sa vrednosti 17.5 na osnovu toga što ova tačka generiše podinterval koji je sledeći podinterval sa najmanjim brojem objekata (4. atribut pripada reduktu). Za ovakvu redukciju dobija se veći redukt {a1, a2, a3, a4}, veći broj pravila i lošiji rezultat klasifikacije. Ukupni rezultat matrice konfuzije je 0.95 što je lošije u odnosu na rezultat bez redukcije tačaka reza.

Može se zaključiti da se kod već odličnog rezultata klasifikacije, dobijenog pravilima na osnovu redukta (dobijenog nad podacima diskretizovanim algoritmom maksimalne razberivosti), naknadnom redukcijom tačaka reza nije dobio bolji rezultat.

5.3.2 Redukcija tačaka reza dobijenih algoritmom maksimalne razberivosti nad bazom sa lošim rezultatom klasifikacije – prikaz segmentacije multimodal raspodele

Rezultat klasifikacije baze Blood Transfusion Service Center na osnovu podataka diskretizovanih algoritmom maksimalne razberivosti je 0.34 (poglavlje 5.2). Ako se posmatraju tačke reza (Dodaci 2) i dobijen redukt {a1, a2, a4}, može se uočiti da je 4. atribut sa vrlo nepravilnom multimodal raspodelom i da ima veliki broj tačaka reza. Sa

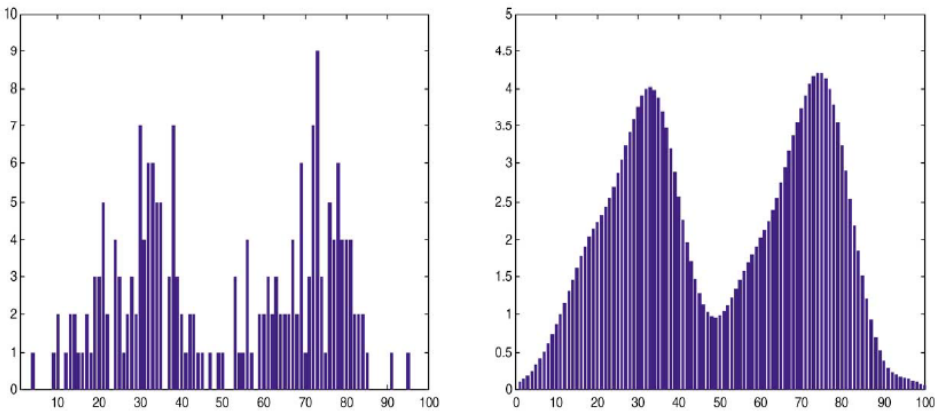
histograma 4. atributa se može primetiti da njegove tačke reza prate lokalne maksimume ili normalnu raspodelu oko lokalnog maksimuma, kao što je zaključeno u poglavlju 5.1. Da bi se bolje prikazala ideja još jednom će se prikazati histogram 4. atributa baze Blood Transfusion Service Center sa tačkama reza na osnovu disktretizacije algoritmom maksimalne razberivosti sa preciznijom rezolucijom (slika 43):



Slika 43. Raspodela podataka 4. atributa T (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

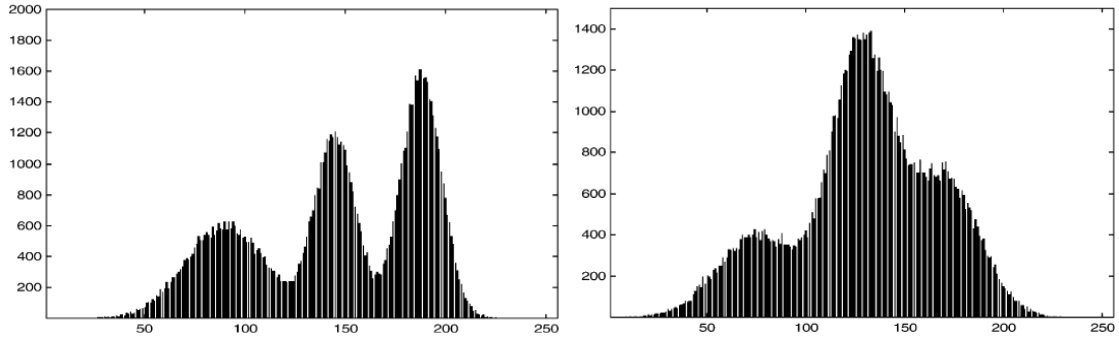
Zaključci poglavlja 5.1 vezani za položaj tačaka reza na histogramima iz prve grupe, naveli su na ideju deljenja - segmentacije histograma koji pripadaju drugoj grupi. Pre nego što se nastavi analiza raspodele 4. atributa baze Blood Transfusion Service Center, pokazaće se načini segmentacije multimodal raspodele.

Istraživanjem segmentacija i dekompozicije multimodal raspodela, primećena je slična ideja naročito kod istraživača iz oblasti Računarske grafike. Postoji nekoliko radova iz ove oblasti [Chang i dr., 2002; Delon i dr., 2007, Chaudhuri, 2010]. U radu [Chang i dr., 2002] opisana je dekompozicija multimodal raspodele tako što se postojeći histogram najpre „ugladi“ – tako da funkcija raspodele predstavlja glatku krivu kao što je pokazano na slici .44 (slika je preuzeta iz rada [Chang i dr., 2002]). Na osnovu preseka glatkih krivih (ili određivanjem minimuma) dobija se tačka prag (*treshold*) koja deli histogram na dva dela.



Slika 44. Originalni histogram (levo) i glatki histogram (desno)

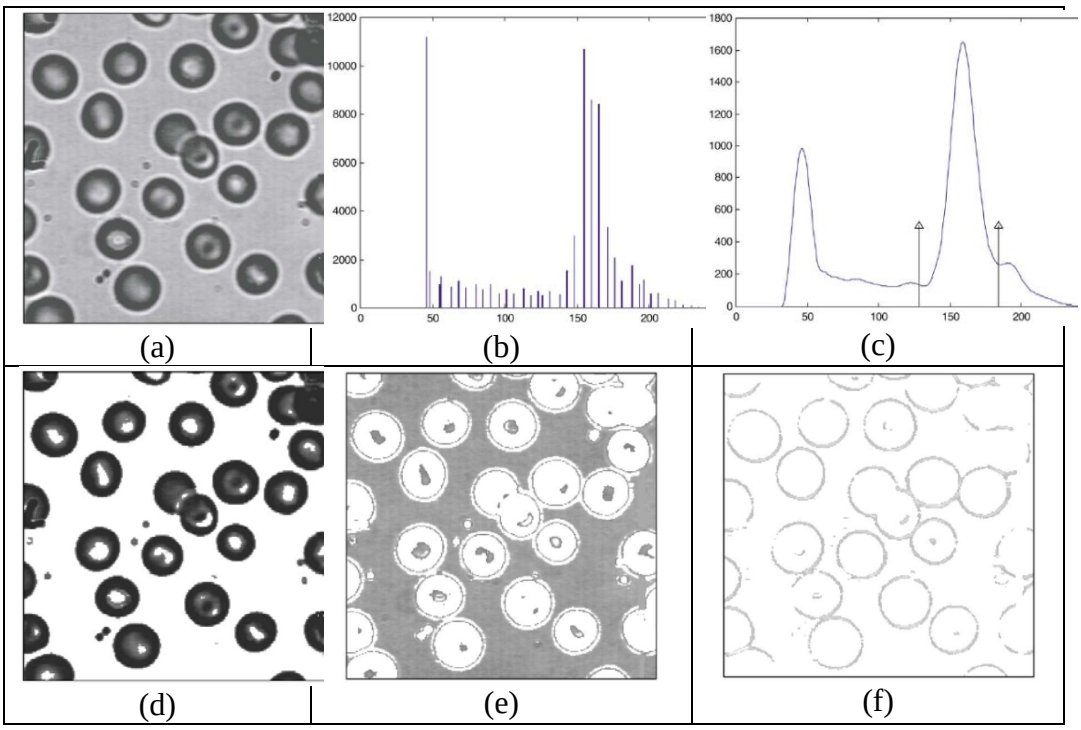
U odnosu na to koliko su glatke krive međusobno razmaknute, moguće je razlikovati izrazito razmaknute i neizrazito razmaknute histograme kao što je prikazano na slici 45 (slika je preuzeta iz rada [Chang i dr., 2002]): bez obzira koliko su razmaknuti imaju isti broj pragova, tačaka na osnovu kojih se obe dele na tri dela.



Slika 45. Tri izrazito razmaknuta histograma (levo) i tri neizrazito razmaknuta histograma (desno)

Primer 5.3.2.1.

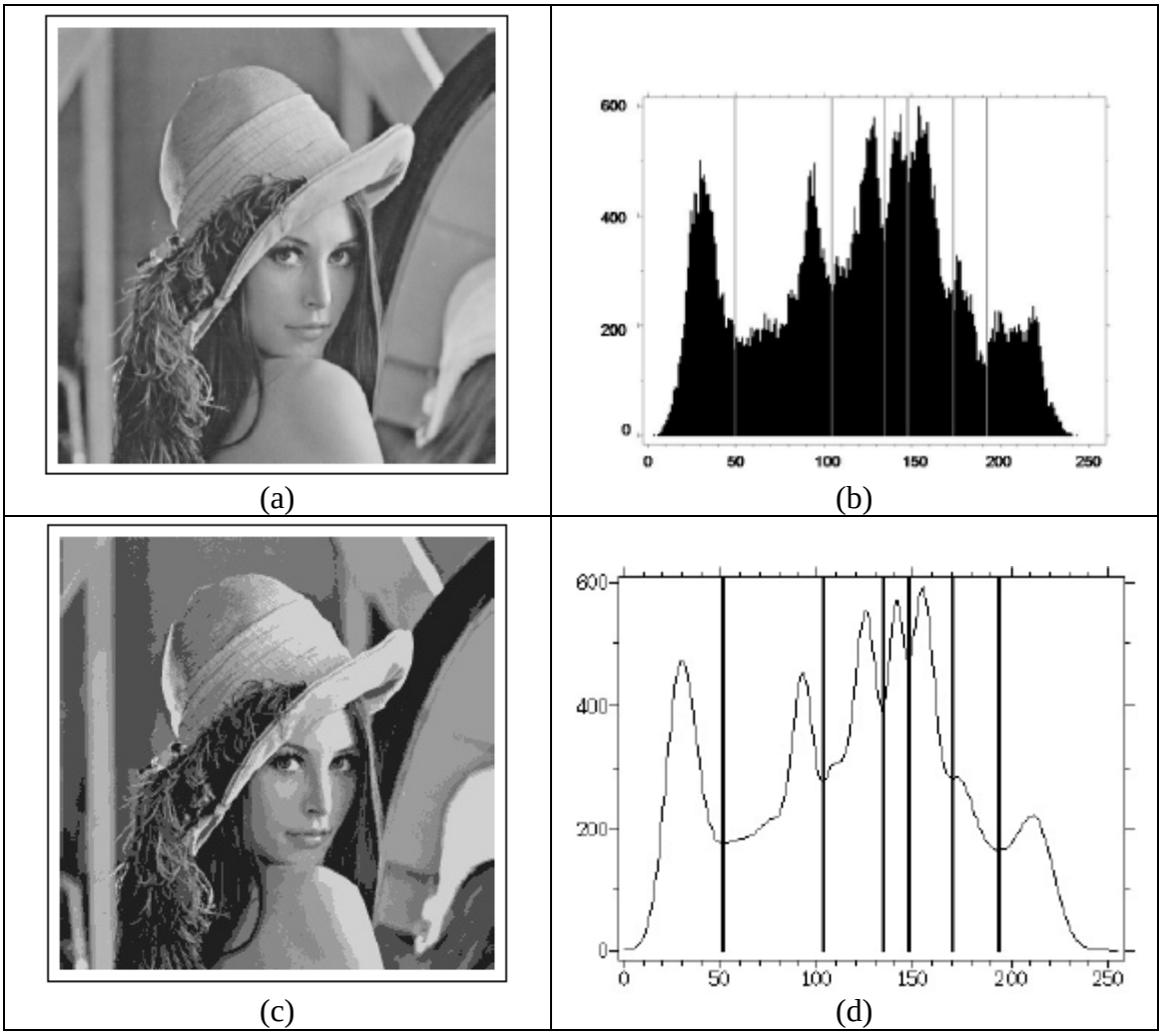
Kao primer koji je urađen u grafici, navodi se analiza slike krvi, čiji histogram ima dva praga. Na slici 46 dat je prikaz slike krvi, histogram, glatka kriva dobijena histogramom i slike dobijene na osnovu segmentacije histograma (primer i slike su preuzete iz rada [Chang i dr., 2002]). Na osnovu prvog dela segmentacije iz intervala (0, 128) dobijena je slika ćelija krvi, na osnovu drugog dela segmentacije iz intervala (128, 184) dobijena je slika krvne plazme i na osnovu trećeg dela segmentacije iz intervala (184, 250) dobijena je slika membrane ćelija.



Slika 46. Slika krvi (a), histogram (b), glatki histogram (c) i rezultati dobijeni na osnovu segmentacije histograma, ćelije krvi (d), krvna plazma (e) i membrane ćelija (f)

Primer 5.3.2.2.

Za razliku od primera 5.3.2.1 gde je segmentacija rađena na osnovu glatkih krivi, segmentaciju histograma je moguće raditi na osnovu lokalnih minimuma. U radu [Delon i dr., 2007] prikazan je algoritam Fine to Coarse (FTC) Segmentation Algorithm, segmentacije histograma na osnovu definisanja liste lokalnih minimuma. On radi objedinjavanje intervala u skladu sa zadovoljenjem definisane unimodal hipoteze. Na slici 47 prikazana je fotografija, njen histogram i fotografija na osnovu segmentiranog histograma (slike su preuzete iz rada [Delon i dr., 2007]). Histogram (slika 47 – deo (b)) je na početku inicijalizacije FTC algoritma imao 60 lokalnih minimuma u okviru 256 početnih tačaka reza. Rezultat dobijene segmentacije je skoro istovetan segmentacijom na osnovu glatkih krivi histograma (slika 47 – deo (d)).



Slika 47. (a) Slika (256 × 256 pixels) Lena, (b) histogram slike pod (a), slika Lena kvantizirana na osnovu 7 nivoa segmentacije histograma na osnovu (b), (d) segmentacija glatkog histograma

Time je potvrđena mogućnost dobijanja segmentacije histograma na osnovu objedinjavanja manjih delova histograma u skladu sa poštovanjem prepoznavanja unimodal raspodele.

Istraživači iz Francuske sa Univerziteta Paris Descartes razvili su ideju segmentacije histograma na čitavu kolor paletu, odnosno na više histograma istovremeno [Delon i dr., 2005 a; Delon i dr., 2005 b; Delon i dr., 2007; Guillemot i dr., 2016] u cilju povezivanja bitnih elemenata kolor slike.

U okviru rada sa dikretizacijama ideja segmentacije multimodal raspodele je nastala u okviru primećivanja zakonitosti položaja tačaka reza kod diskretizacije algoritmom maksimalne razberivosti. Razrada te ideje u okviru objedinjavanja manjih delova histograma kod diskretizacije algoritmom maksimalne razberivosti, kao i objedinjavanje svih multimodal histograma je takođe primećena i opisana u radu [Ognjenović i dr., 2016]. U ovoj disertaciji ideja je primenjena tako da povezuje sve multimodal histograme a takođe i histograme ostalih atributa iz prve grupe.

Na osnovu prethodno analiziranih algoritama i primera segmentacije multimodal raspodele koji su slični ideji koja će se izneti, uradiće se redukcija tačaka reza baze Blood Transfusion Service dobijenih algoritmom maksimalne razberivosti, na dva načina.

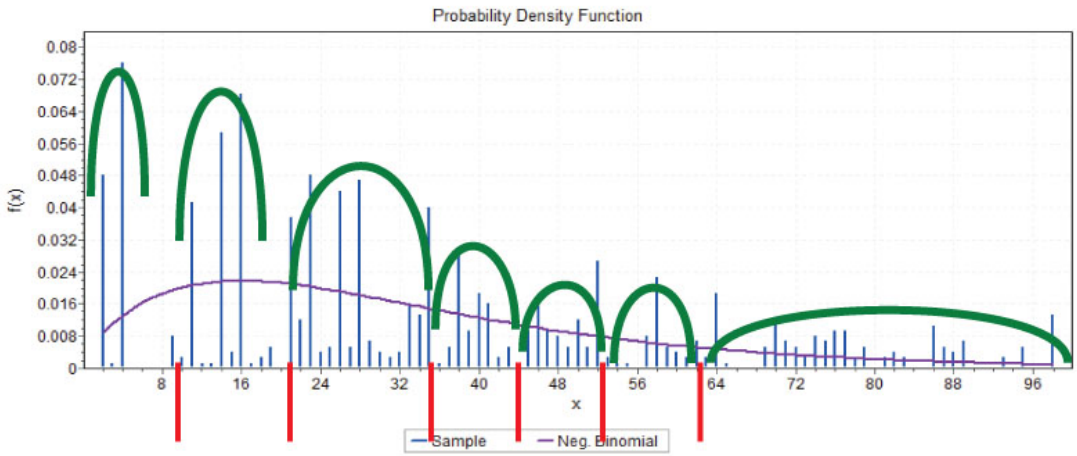
Prvi način.

U odnosu na sve histograme baze uradiće se:

I SEGMENTACIJA MULTIMODAL RASPODELA U MERI POSTOJANJA TAČAKA REZA. TO ZNAČI DA SE OD SVIH TAČAKA REZA BIRAJU ONE TAČKE KOJE SU NAJBLIŽE TAČKAMA PRAGOVIMA SEGMENTACIJE. TAKVE TAČKE SU NAZVANE FIKSNE TAČKE.

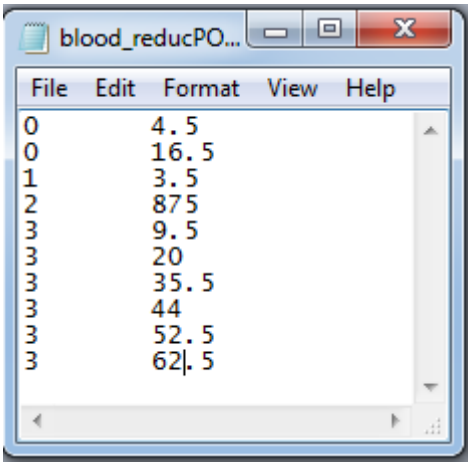
II BRISANJE SVIH TAČAKA REZA OSIM FIKSNIH TAČAKA (UKLJUČUJUĆI I TAČKE KANDIDATE REZA ZA NEDISKRETIZOVANE ATRIBUTE).

U odnosu na histogram četvrtog atributa sa tačkama reza koji je prikazan na slici 43., fiksne tačke i odgovarajuća segmentacija histograma je prikazana na slici 48. Fiksne tačke su nacrtane vertikalnim linijama na apcisi.



Slika 48. Raspodela podataka 4. atributa T (baze Blood Transfusion Service Center) sa fiksnim tačkama i odgovarajućom segmentacijom histograma

Ako se nad bazom Blood Transfusion Service Center, istim postupkom redukuju tačke reza i kod ostalih atributa, onda se za tako redukovane tačke reza koje su ručno unete u sistem Rosetta (slika 49), dobija rezultat klasifikacije prikazan na slici 50.



Slika 49. Redukovan skup tačaka reza baze Blood Transfusion Service Center na osnovu prvog načina

		Predicted			
		0	1	Undefined	
Actual	0	235	46	5	0.821678
	1	59	29	0	0.329545
	Undefined	0	0	0	Undefined
		0.79932	0.386667	0.0	0.705882

Slika 50. Rezultat klasifikacije baze Blood Transfusion Service Center na osnovu redukcije tačaka reza na prvi način

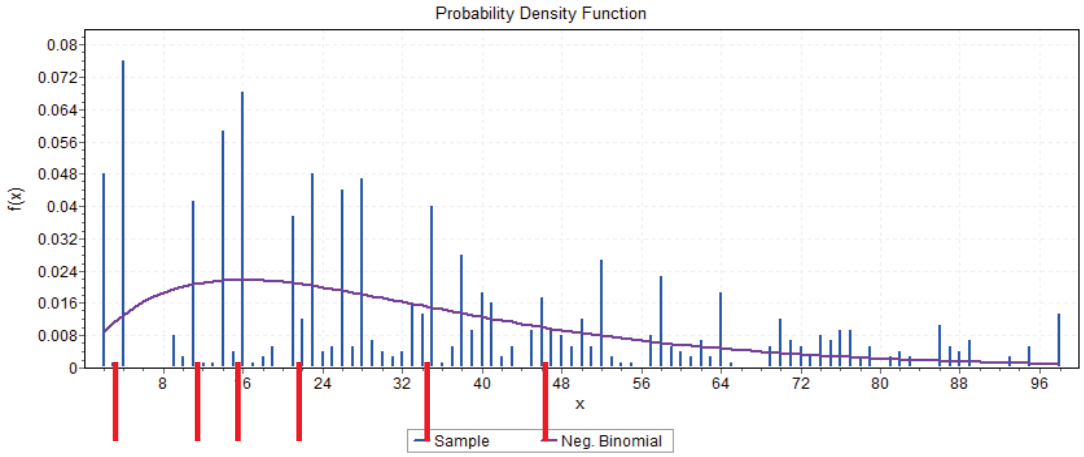
Na osnovu rezultata matrice konfuzije može se videti da se na ovaj način diskretizovana tabela bolje klasifikovala na osnovu ukupnog rezultata. Problem koji je evidentan je da se pored značajnog povećanja broja objekata koji se pravilno klasifikuju, povećao i broj objekata koji se nepravilno klasifikuju. Pored posmatranja matrica konfuzije, ako se posmatraju i pravila na osnovu kojih je izvršena klasifikacija, može se uočiti sledeće:

- na osnovu diskretizacije algoritmom maksimalne razberivosti dobijeno je 265 pravila od kojih je 12 nepreciznih – procenat nepreciznih pravila je 5%;
- na osnovu redukcije tačaka reza na prvi način dobijeno je 33 pravila od kojih je 20 nepreciznih – procenat nepreciznih pravila se povećao na 61%.

Povećanjem broja pravila koja imaju operator OR u THEN delu pravila, smanjuje se razberivost. Zbog toga matrica konfuzije može da ima dobar rezultat a da se u stvari ne zna tačna odluka klasifikacije za konkretan objekat.

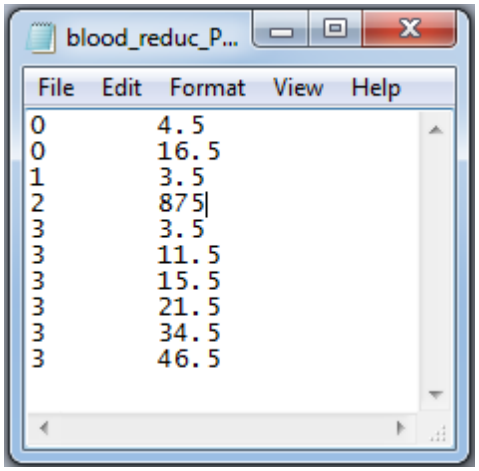
Drugi način.

Da bi se pokazala značajnost odabira fiksnih tačaka u odnosu na neku drugu redukciju tačaka, u okviru baze Blood Transfusion Service Center će se ponovo poći od tačaka reza dobijenih algoritmom maksimalne razberivosti. Izuzev četvrtog atributa, kod ostalih atributa uzeće se fiksne tačke kao u Prvom načinu. Samo kod četvrtog atributa namerno će se zaobići fiksne tačke, odnosno izabraće se tačke koje se nalaze unutar segmentacija, a fiksne tačke će se izbrisati. Izbor tačaka unutar segmentacija je takav da određuju približan broj objekata u odnosu na broj objekata koji su određivale fiksne tačke. Takav izbor tačaka reza četvrtog atributa je prikazan na histogramu na slici 51, i u odnosu na prvi način, ovo je jedina izmena.



Slika 51. Izmenjen skup tačaka reza kod 4. atributa T (baze Blood Transfusion Service Center) – umesto fiksnih proizvoljne tačke

Za ovako izmenjene tačke reza četvrtog atributa, koje su ručno unete u sistem Rosetta (slika 52), dobija se rezultat klasifikacije prikazan na slici 53.



Slika 52. Izmenjen skup tačaka reza za četvrti atribut baze Blood Transfusion Service - umesto fiksnih proizvoljne tačke

		Predicted			
Actual		0	1	Undefined	
	0	259	25	2	0.905594
	1	63	25	0	0.284091
	Undefined	0	0	0	Undefined
		0.804348	0.5	0.0	0.759358

Slika 53. Rezultat klasifikacije za izmenjen skup tačaka reza za četvrti atribut baze Blood Transfusion Service - umesto fiksnih proizvoljne tačke

Ono što deluje kao napredak u poslednjoj matrici konfuzije, rezultat klasifikacije se sa 71% povećao na 76%, je u stvari problem jer je klasifikacija dobijena povećanjem broja pravila koja u THEN delu imaju operator OR. Na bazi diskretizacije izmenjenog skupa tačaka reza sa slike 52, dobijeno je ukupno 29 pravila, od kojih njih 19 ima operator OR u THEN delu – procenat nepreciznih pravila se povećao na 66%.

U odnosu na redukciju tačaka reza iz prvog primera, za četvrti atribut je uzet isti broj tačaka reza kao u prvom načinu i to tako da određuju približno jednak broj objekata u odnosu na broj objekata koji su određivale fiksne tačke. Upoređujući prvi i drugi način može se primetiti sledeće:

- prvim načinom dobijeno je 33 pravila od kojih je 20 nepreciznih – procenat nepreciznih pravila bio je 61%;
- drugim načinom dobijeno je 29 pravila od kojih je 19 nepreciznih – procenat nepreciznih pravila se povećao na 66%.

Nad nekim drugim kombinacijama tačaka reza, pod uslovom istog broja tačaka reza za svaki atribut (isti broj kao u Prvom i Drugom načinu) takođe dolazi do povećanja procenta neprecinih pravila a ukupni rezultat klasifikacije varira, odnosno za neke kombinacije je i manji od rezultata dobijenog Prvim načinom.

6 ALGORITAM APROKSIMATIVNE DISKRETIZACIJE MAKSIMALNE RAZBERIVOSTI APPROX MD

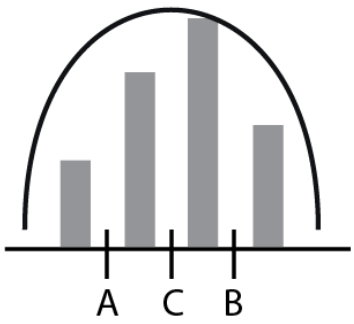
Na osnovu istraživanja o uticaju raspodele podataka na algoritme diskretizacije, analizu klasifikacije na osnovu raspodele podataka nad deset baza, na osnovu ideje o redukciji tačaka reza iz poglavlja 5, definisani su uslovi redukcije tačaka reza i na osnovu njega konstruisan je algoritam aproksimativne diskretizacije maksimalne razberivosti - APPROX MD algoritam u odnosu na raspodele atributa, redukt i konzistentnost tabelarno organizovane baze.

6.1 USLOVI REDUKCIJE TAČAKA REZA

Na osnovu istraživanja urađenog u poglavlju 5, na osnovu uticaja raspodele podataka na algoritam diskretizacije maksimalne razberivosti, na osnovu uticaja raspodele podataka na kvalitet klasifikacije, na osnovu ideje redukcije tačaka reza dobijenih algoritmom maksimalne razberivosti i njenim uticajem na redukt, pravila i klasifikaciju, definisani su uslovi redukcije tačaka reza dobijenih algoritmom maksimalne razberivosti:

1. da li tačka reza pripada ili ne pripada atributu koji je element redukta – ako tačka reza pripada atributu koji ne pripada reduktu njena redukcija je moguća samo uz tačku reza nekog drugog atributa koji pripada reduktu;
2. kod 1. grupe raspodela redukcija tačke reza se radi na osnovu toga da li generiše podinterval sa najmanjim brojem objekata (u odnosu na sve podintervale svih atributa) – ako se podinterval sa najmanjim brojem objekata ne nalazi na kraju intervala (domena atributa) onda se od dve tačke reza koje određuju podinterval sa najmanjim brojem objekata bira ona koja ima manji broj objekata u oba podintervala koja određuje;
3. kod 2. grupe raspodela redukcija tačaka reza se radi nakon definisanja fiksnih tačaka u odnosu na već dobijene tačke reza – nakon segmentacije raspodele, redukcija tačaka reza se radi u okviru intervala segmentacije kao kod 2. uslova.

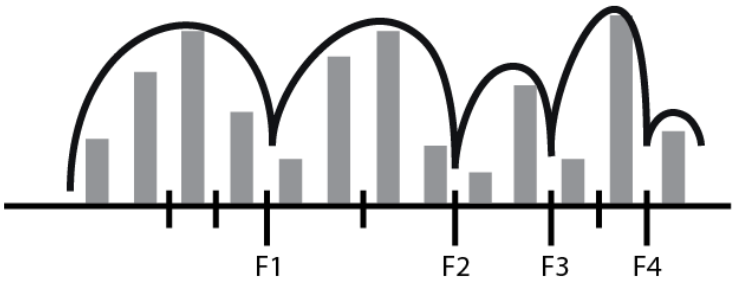
U okviru drugog uslova, na primeru je pokazan uticaj histograma u određivanju podintervala sa najmanjim brojem objekata. Na slici 54 dat je histogram nekog atributa i tačke reza. Prva tačka reza za redukciju (izbacivanje) bi bila tačka reza A. Druga bi bila tačka reza B i treća po redu bi bila tačka reza C. Izbacivanjem svih tačaka došlo bi se do redukcije atributa iz procesa diskretizacije.



Slika 54. Primer histograma sa tačkama reza

Redukcija početnih tačaka reza ovakvog tipa je veoma čest korak u raznim algoritmima diskretizacije i predstavlja jednostavan način objedinjavanja podintervala.

U okviru trećeg uslova definisanje segmentacije histograma u odnosu na već dobijene tačke reza predstavlja cepanje multimodal raspodelu na delove koji predstavljaju raspodelu iz 1. grupe raspodela. Ako se posmatra histogram na slici 55 sa tačkama reza, onda bi tačke reza označene slovima F1, F2, F3 i F4 bile uzete kao fiksne. Nad ostalim tačkama reza bi se radila redukcija u okviru svakog podintervala kao kod 2. uslova. U okviru rada sa histogramima pokazano je da se kod multimodal raspodele dobija veći broj tačaka reza nego kod ostalih raspodela. U slučaju redukcije neke od tačaka F1, F2, F3 i F4, drastično bi se povećao broj nepreciznih pravila jer bi sa jedne strane veći broj atributa postao deo redukta, dok bi sa druge strane mali broj intervala imao kao rezultat iste IF delove i raličite THEN delove pravila. Ukupan rezultat bi se „popravio“ na drastičnu štetu konzistentnosti diskretizovane tabele.



Slika 55. Primer histograma sa fiksnim tačkama reza

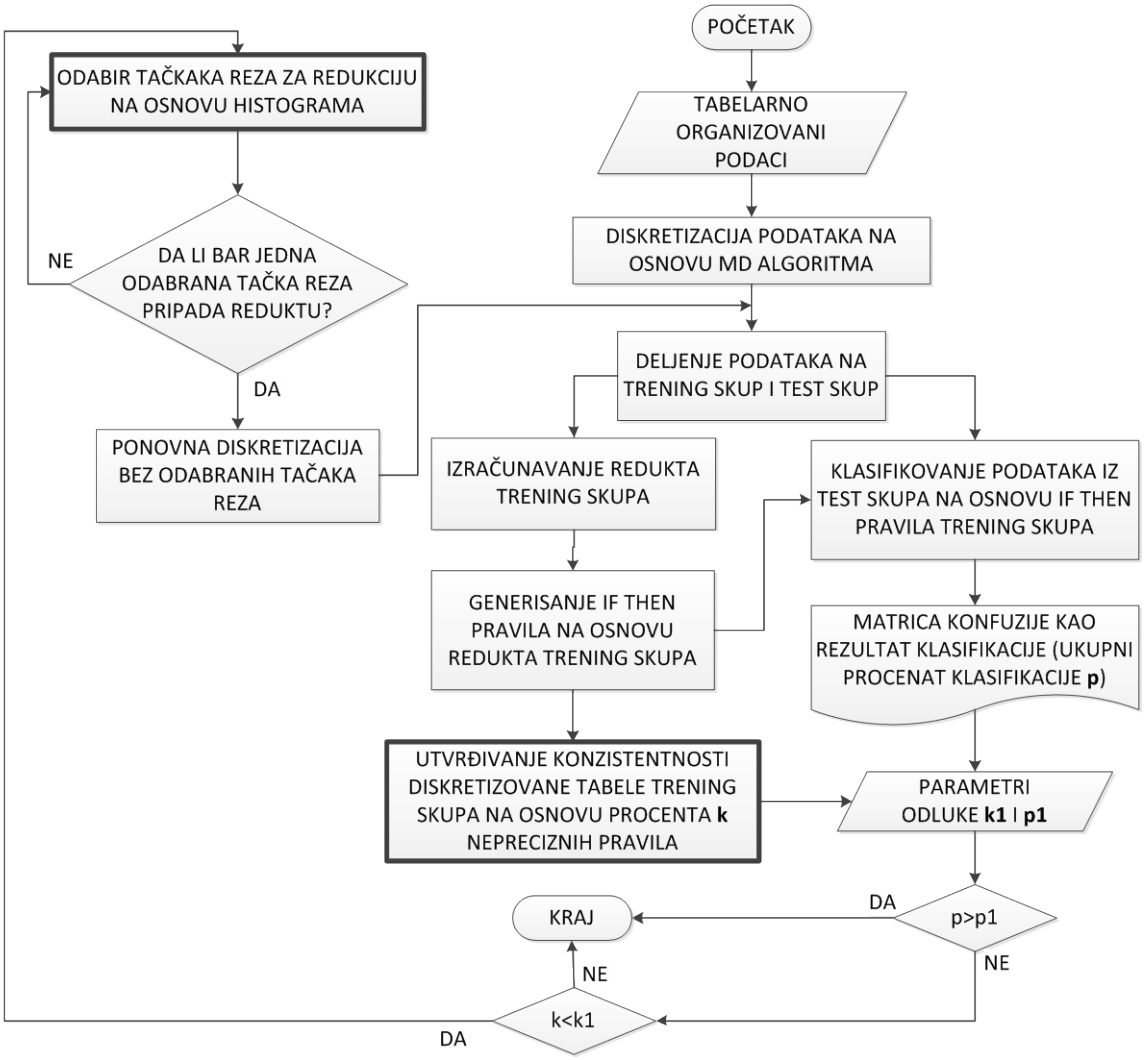
Na osnovu svega razmatranog napravljen je algoritam koji je primenljiv na tačke reza dobijene algoritmom maksimalne razberivosti, koji prati opisane uslove redukcije tačaka reza predložene u ovom poglavlju i koji je testiran na izabranim bazama.

6.2 KONSTRUKCIJA ALGORITMA APROKSIMATIVNE DISKRETIZACIJE MAKSIMALNE RAZBERIVOSTI APPROX MD

Algoritam aproksimativne diskretizacije APPROX MD je prikazan blok shemom na slici 56. Baziran je na rezultatima – tačkama reza algoritma diskretizacije maksimalne razberivosti MD i na parametrima vezanim za procenat nepreciznih pravila, ukupni procenat klasifikacije i broju tačaka redukcije. Dva procesa su istaknuta kao bitna i to:

I. Odabir tačaka reza za redukciju na osnovu histograma i

II. Utvrđivanje konzistentnosti diskretizovane tabele trening skupa na osnovu procenta k nepreciznih pravila.



Slika 56. Blok shema algoritma aproksimativne diskretizacije maksimalne razberivosti APPROX MD

Drugi proces je trivijalan, tako da je samo prvi proces posebno definisan i prikazan. Odabir tačaka reza za redukciju na osnovu histograma, na osnovu uslova sumiranih u poglavlju 6.1 moguće je uraditi na osnovu:

1. određivanja fiksnih tačaka,
2. broja tačaka redukcije.

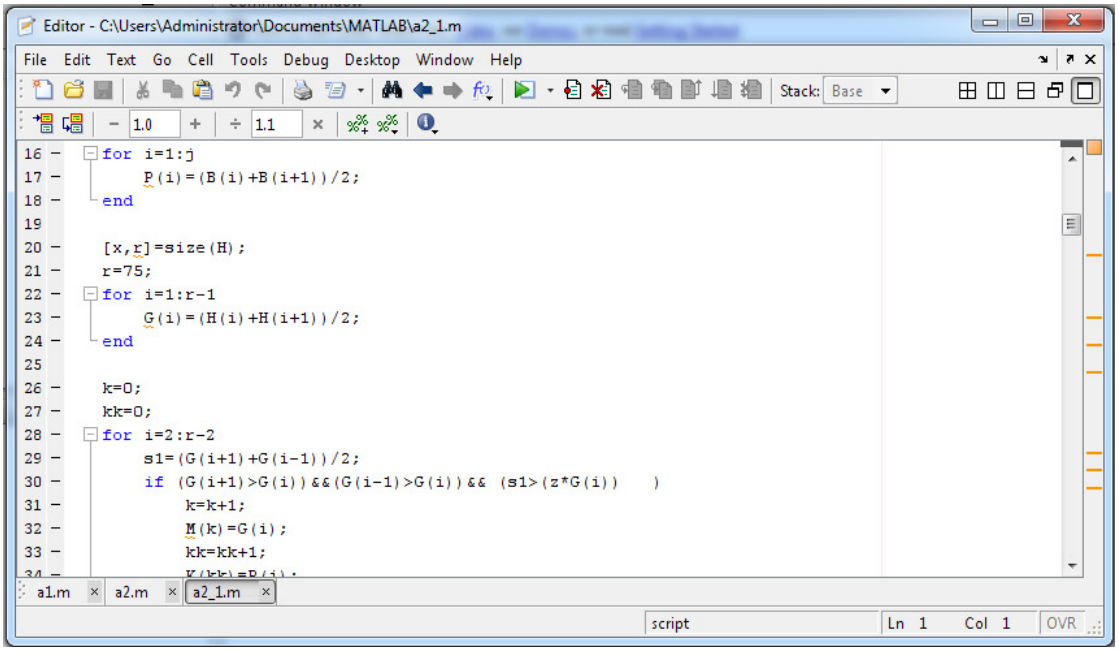
Određivanje fiksnih tačaka je moguće od strane eksperta ili automatski. Za automatsko određivanje fiksnih tačaka urađen je algoritam FixedPoints

6.2.1 Algoritam FixedPoints

Za histogram atributa sa n bar-ova, definisane su tačke b_1, b_2, …, b_{max} koje dele apcisu histograma u odnosu na bar-ove: svaki bar histograma se nalazi između dve susedne tačke b_k i b_{k+1}. Dovoljno je uraditi grubo "uglačavanje" histograma, tako da se tačkama p_k koje se nalaze između tačaka b_{k-1} i b_k pridružuju novi bar-ovi g_k koji su kandidati za ispunjavanje logičkog uslova postojanja segmentacije histograma. Barovi g_k se ne dobijaju na osnovu podataka baze, već na osnovu srednje vrednosti početnih susednih barova: $g_k = \frac{b_{k-1} + b_k}{2}$. Procena i osmišljavanje grubog „uglačavanja" je rađeno na osnovu analize međusobnih odnosa visine bar-ova elementarnom geometrijom [Matlab – Histogram, 2015] [Cameron, 2015].

Nakon grubog „uglačavanja" definisan je kriterijum segmentacije tako modifikovane multimodal raspodele. U odnosu na postojanje lokalnog minimuma g_i, radi se upoređivanje bar-ova $\frac{g_{i-1} + g_{i+1}}{2}$ i g_i. U slučaju da je bar $\frac{g_{i-1} + g_{i+1}}{2}$ značajno veći od g_i, tačka g_i postaje okolina fiksne tačke, odnosno njoj najbliža tačka reza postaje fiksna tačka. Ako bar $\frac{g_{i-1} + g_{i+1}}{2}$ nije značajno veći od g_i, radi se upoređivanje bar-ova $\frac{g_{i-2} + g_{i+1}}{2}$ i g_i, pod uslovom da je $g_{i-2} > g_{i-1}$, pa upoređivanje bar-ova $\frac{g_{i-1} + g_{i+2}}{2}$ i g_i, i tako dalje do zadovoljenja uslova.

U Matlab-u je urađena implementacija ove ideje, na osnovu koje je moguće odrediti fiksne tačke za definisane uslove (slika 57).



Slika 57. Određivanje fiksnih tačaka u Matlab-u

Grubo „uglačavanje" histograma je dovoljno zbog toga što za određivanje fiksnih tačaka nije neophodno da se tačno odredi segmentacija multimodal raspodele, već samo približna vrednost na osnovu koje se određuje njoj najbliža tačka reza.

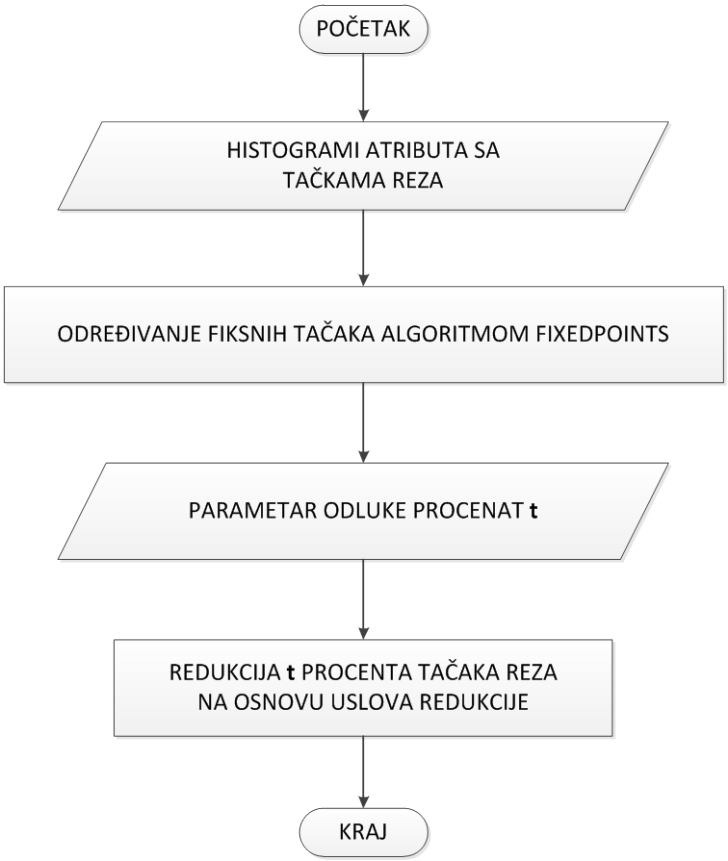
U okviru ove disertacije nije predviđena implementacija algoritma aproksimativne diskretizacije APPROX MD, tako da će se razrada i kompletna implementacija algoritma FixedPoints raditi u budućim istraživanjima u okviru implementacije algoritma APPROX MD.

6.2.2 Odabir tačaka reza za redukciju na osnovu histograma

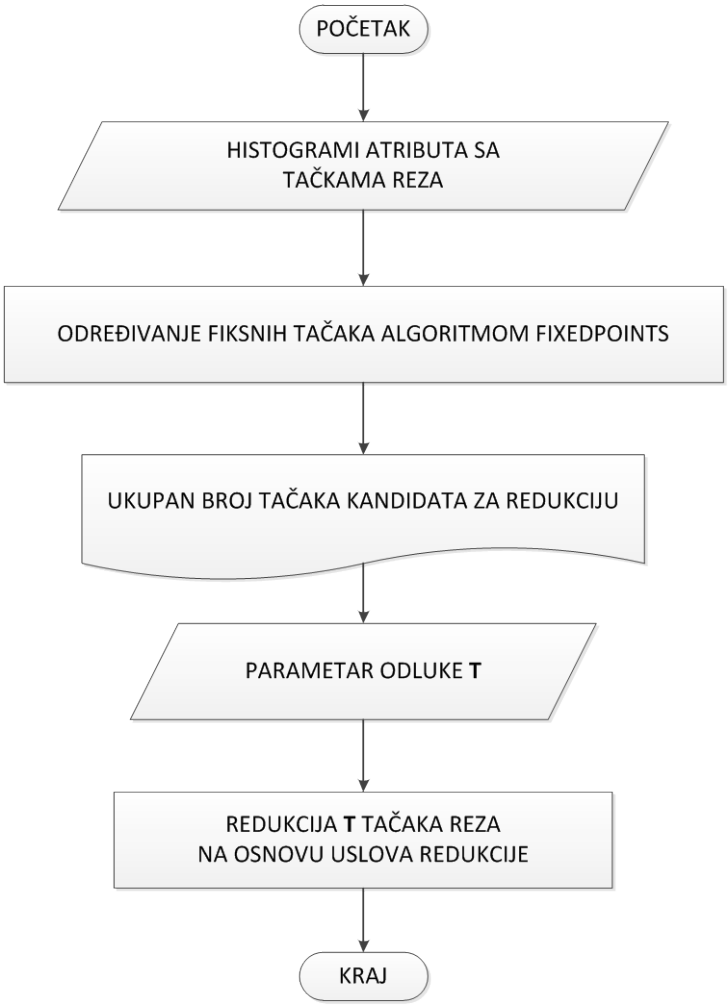
Odabir tačaka reza za redukciju se radi u odnosu na fiksne tačke. Parametrom se definiše broj tačaka reza koje će se izbrisati u svakoj iteraciji. Redukcija tačaka reza se radi u skladu sa uslovima redukcije tačaka reza (poglavlje 6.1). Unos parametara moguć je na dva načina:

I. unosom procenta redukcije od ukupnog broja tačaka reza **t**, i u tom slučaju bi se u svakoj iteraciji radila redukcija istog procenta od preostalog ukupnog broja tačaka reza (slika 58);

II. od strane eksperta, unosom broja tačaka reza za redukciju **T** na osnovu analize dokumenta Fiksne tačke i ostale tačke reza (slika 59) – u tom slučaju je moguće da se u svakoj iteraciji radi redukcija različitog broja tačaka.



Slika 58. Blok shema procesa Odabir tačaka reza za redukciju na osnovu histograma (za unos procenta redukcije od ukupnog broja tačaka reza)



Slika 59. Blok shema procesa Odabir tačaka reza za redukciju na osnovu histograma (za unosa broja T tačaka reza za redukciju od strane eksperta)

Ovako definisan algoritam aproksimativne diskretizacije omogućuje da se ukupni rezultat klasifikacije „kontroliše“ na osnovu broja nepreciznih pravila. Time se pored nove klasifikacije dobijaju i novi rezultati za redukt skupa.

6.3 ANALIZA PRIMENE ALGORITMA APPROX MD

U okviru analize primene algoritma APPROX MD izabraće se parametri odluke k1=20% i p1=95%. Odabir p1=95% je urađen zbog kasnije komparacije sa algoritmom maksimalne razberivosti za aproksimatvna rešenja za α=0,95. Za određivanje tačaka reza za brisanje (redukciju), koristiće se verzija za eksperta, odnosno za svaku iteraciju će se navesti broj tačaka brisanja. Za analizu primene algoritma APPROX MD uzeće se dve baze i to baza Iris koja ima odličan rezultat klasifikacije i baze Blood Transfusion Service Center koja ima loš rezultat klasifikacije. Obe baze imaju raspodele iz obe grupe raspodela.

6.3.1 Primena algoritma APPROX MD nad bazom koja ima dobar rezultat klasifikacije

Matrica konfuzije baze Iris ima ukupne rezultate k=0% i p=96%, tako da primena algoritma APPROX MD za zadate parametre odluke k1=20% i p1=95% neće proizvesti redukciju tačaka reza. U prvoj iteraciji će uslov p>p1 biti zadovoljen i algoritam će prestati sa radom. To znači da se primenom algoritma APPROX MD dobijaju isti rezultati kao i bez njegove primene.

6.3.2 Primena algoritma APPROX MD nad bazom sa lošim rezultatom klasifikacije

Matrica konfuzije baze Blood Transfusion Service Center ima ukupne rezultate k=4% i p=34%, tako da će primena algoritma APPROX MD za zadate parametre odluke k1=20% i p1=95% proizvesti redukciju tačaka reza.

U prvoj iteraciji će zbog toga što uslov p>p1 nije zadovoljen (34% nije veće od 95%), biti izvršena provera k<k1. Kako je 4% < 20% (uslov je zadovoljen), radiće se odabir tačaka reza za redukciju na osnovu histograma podataka. Ukupno postoji 53 tačaka reza generisanih za diskretizovane atribute (Dodaci 2, slika 71) i tačke reza jednog nediskretizovanog atributa, 3. atributa M (Dodaci 2, slika 74) kojih ima 31. Od toga je 10 fiksnih tačaka (Poglavlje 5.3.2, slika 49). Broj tačaka kandidata za redukciju iznosi (53+31)-10=74. Ako se izabere 20 tačaka za redukciju na osnovu uslova za redukciju (od izabranih tačaka reza za redukciju ima onih koje pripadaju atributima koji su deo redukta), onda se dobijaju vrednosti k=5% i p=34%. Procenat konzistentnost se zanemarljivo povećao, a rezultat klasifikacije je ostao isti.

U drugoj iteraciji uslov p>p1 ponovo nije zadovoljen (34% nije veće od 95%), pa će biti izvršena provera k<k1. Kako je 5% < 20% (uslov je zadovoljen), radiće se odabir tačaka reza za redukciju na osnovu histograma podataka. Tačaka kandidata za brisanje ima 54. Ako se ponovo izabere redukcija 20 tačaka reza na osnovu uslova za redukciju (od izabranih tačaka reza za redukciju ima onih koje pripadaju atributima koji su deo redukta), onda se dobijaju nove vrednosti k=12% i p=47%.

U trećoj iteraciji uslov p>p1 ponovo nije zadovoljen (47% nije veće od 95%), pa će biti izvršena provera k<k1. Kako je 12% < 20% (uslov je zadovoljen), radiće se odabir tačaka reza za redukciju na osnovu histograma podataka. Tačaka kandidata za brisanje ima 34. Ako se ponovo izabere brisanje 20 tačaka reza na osnovu uslova za redukciju

(od izabranih tačaka reza za redukciju ima onih koje pripadaju atributima koji su deo redukta), onda se dobijaju nove vrednosti k=53% i p=72%.

U četvrtoj iteraciji uslov p>p1 ponovo nije zadovoljen (027% nije veće od 95%), pa će biti izvršena provera k<k1. Kako 53% nije manje od 20% (uslov nije zadovoljen), algoritam će završiti sa radom. Primenom algoritma APPROX MD dobio se rezultat k=53% i p=72%.

7 KOMPARACIJA ALGORITMA APPROX MD SA ALGORITMOM MD I ALGORITMOM MD SA APROKSIMATIVNIM REŠENJIMA ZA α=0,95

Nad bazama koje su korišćene u poglavljima 5 i 6, primenjen je algoritmom diskretizacije maksimalne razberivosti za aproksimatvna rešenja sa vrednošću α=0,95. U tabeli XVIII prikazani su uporedni rezultati procenta nepreciznih pravila **k** i ukupnog procenta klasifikacije **p** iz matrice konfuzije, dobijeni sledećim algoritmima:

- algoritmom diskretizacije maksimalne razberivosti MD
- algoritmom diskretizacije MD sa aproksimativnim rešenjem za α=0,95
- algoritmom aproksimativne diskretizacije APPROX MD za parametre k1=20% i p1=95%.

Naziv baze	Rezultati dobijeni algoritmom diskretizacije MD	Rezultati dobijeni algoritmom MD sa aproksimativnim rešenjima za α=0,95	Rezultati dobijeni aproksimativnim algoritmom APPROX MD
1. Iris	k=0% p=96%	k=50% p=63%	k=0% p=96%
2. Blood Transfusion Service Center	k=5% p=34%	k=38% p=71%	k=53% p=72%
3. Banknote Authentication	k=0% p=97%	k=75% p=91%	k=0% p=97%
4. Glass Identification	k=0% p=34%	k=0% p=8%	k=43% p=51%
5. Wilt Data Set	k=0% p=96%	k=0% p=1%	k=0% p=96%
6. Breast Cancer Wisconsin Data Set	k=0% p=77%	k=0% p=57%	k=22% p=92%
7. Cardiotocography	k=0% p=39%	k=0% p=7%	k=0% p=78%
8. Statlog (Australian Credit Approval)	k=0% p=8%	k=0% p=0,3%	k=28% p=29%
9. Haberman's Survival Data Set	k=0% p=15%	k=67% p=71%	k=32% p=56%
10. Challenger USA Space Shuttle O-Ring Data Set	k=0% p=67%	k=25% p=91%	k=25% p=91%

TABELA XVIII. UPOREDNI REZULTATI PROCENTA NEPRECIZNIH PRAVILA **K** I UKUPNOG PROCENTA KLASIFIKACIJE **P**

Rezultati algoritma APPROX MD dobijeni su iteracijama u kojima se brisalo više tačaka u okviru jedne iteracije. U slučaju brisanja jedne ili manjeg broja tačaka u okviru

jedne iteracije, dobijeni rezultat procenat nepreciznih pravila (broj k) ne bi bio mnogo veći od zadatog praramentra k1.

Na osnovu tabele XVIII može se zaključiti da kod algoritma APPROX MD postoji bolji ili, u slučajevima vrlo malih baza jednak rezultat klasifikacije u odnosu na rezultate dobijene algoritmom MD sa aproksimativnim rešenjima za α=0,95. Takođe postoji jednak ili bolji rezultat klasifikacije algoritma APPROX MD u odnosu na algoritam MD, na štetu lošije konzistentnosti diskretizovane tabele.

8 ZAKLJUČCI I SMERNICE BUDUĆIH ISTRAŽIVANJA

Analizom odnosa histograma podataka i algoritama za diskretizaciju pokušana je da se istakne važnost raspodele podataka na konačan rezultat klasifikacije. Na osnovu prikaza položaja tačaka reza na histogramu, istraživanje se prvo fokusiralo na definisanje grupa raspodela bitnih za eksperiment. Dva velika pravca u okviru sistema odlučivanja, stabla odlučivanja i teorija grubih skupova imaju direktnu korist od detaljnih analiza odgovarajućih algoritama diskretizacije. Ovom disertacijom je istraženo područje odnosa osobina podataka datim histogramima i dobijenim tačkama reza određene diskretizacije. Dobijeni rezultati uticaja histograma na tačke reza se mogu koristiti u daljem radu na razvoju algoritama diskretizacije. Cilj definisan ovom disertacijom je ispunjen primenom rezultata istraživanja za konstrukciju algoritma aproksimativne diskretizacije APPROX MD.

8.1 ZAKLJUČCI

U odnosu na definisane konkretne ciljeve, zadatke i hipoteze disertacije na osnovu urađenog rada može se zaključiti sledeće:

Konkretni ciljevi su ispunjeni

Cilj I Urađena je provera odnosa tačaka reza dva algoritma diskretizacije (algoritma diskretizacije maksimalne razberivosti i algoritma diskretizacije baziranog na entropiji) i raspodela podataka. Detaljni opisi i zaključci su prikazani u poglavlju 5.1.

Cilj II Urađena je provera odnosa raspodele podataka u odnosu na rezultat klasifikacije. Detaljni opisi i zaključci su prikazani u poglavlju 5.2.

Cilj III Urađena je provera odnosa rezultata klasifikacije u odnosu na inkozistentnost tabele i algoritam diskretizacije. Detaljni opisi i zaključci su prikazani u poglavlju 5.2.

Cilj IV Proučen je uticaj redukcije broja tačaka reza na inkonzistentnost tabele. Analiza i zaključci su prikazani u poglavlju 5.3.2 i na kraju disertacije u tabeli XVIII.

Tačke reza dobijene algoritmom maksimalne razberivosti kod raspodela iz 1. grupe uglavnom dele interval na podjednak broj podintervala. U slučaju raspodela iz 2. grupe, tačke reza se nalaze u maloj okolini tačaka – pragova segmentacije multimodal histograma. Tačke reza dobijene algoritmom diskretizacije baziranom na entropiji dele interval histograma vrlo nepravilno kod obe grupe raspodela. Naročito u slučaju multimodal raspodela postoji velika disproporcija u broju objekata koje određuju tačke reza.

Za veći broj elemenata redukta i multimodal raspodele uglavnom su se dobijali lošiji rezultati klasifikacije. Kod algoritma maksimalne razberivosti rezultat diskretizacije je generisao diskretizovanu tabelu sa manjom inkonzistentnošću nego algoritam diskretizacije baziran na entropiji. Kod oba algoritma, diskretizacija nad podacima koji u većini pripadaju 1. grupi raspodela rezultovala je manjom inkonzistentnošću

diskretizovane tabele nego kod diskretizacije nad podacima koji u većini pripadaju 2. grupi raspodela.

Redukcija tačaka reza dobijenih algoritmom maksimalne razberivosti predložena je na osnovu uočenih fiksnih tačaka reza koje odgovaraju gruboj segmentaciji multimodal raspodela. Na primeru je pokazan uticaj redukcije fiksnih tačaka na inkonzistentnost diskretizovane tabele u odnosu na redukciju ostalih tačaka reza. Eksperimentom je potvrđeno da redukcija fiksnih tačaka reza više utiče na inkonzistentnost tabele od redukcije ostalih tačaka. Za generisanje fiksnih tačaka od tačaka reza napravljen je algoritam FixedPoins koji je implementiran u Matlab-u.

Zadaci su urađeni

Nad deset izabranih baza urađena je analiza histograma pomoću softvera EasyFit. Primenjeni su predviđeni algoritmi za diskretizaciju podataka: algoritam diskretizacije maksimalne razberivosti i algoritam diskretizacije baziran na entropiji. Klasifikacija je urađena predviđenim softverom sistemom Rosetta korišćenjem Džonsonovog algoritma za izračunavanje redukta. Dat je odgovor na pitanje koliko je rezultat klasifikacije dobar samo na osnovu ukupnog procenta matrice konfuzije a koliko zbog konzistentnosti\inkonzistentnosti tabele.

Ako se u zavisnosti od raspodele podataka u skladu sa definisanim uslovima redukcije brišu samo tačke reza atributa koji nisu deo redukta, redukt ostaje nepromenjen. Ako se briše bar jedna tačka reza atributa koji je deo redukta, onda redukt nad diskretizovanim podacima od preostalih tačaka reza može da se promeni. Algoritam APPROX MD omogućuje da redukcija tačaka reza samo onih atributa koji nisu deo redukta i ne proizvode promenu redukta, a koje zadovoljavaju uslove za redukciju definisane na osnovu histograma, bivaju dopunjene sledećom redukcijom tačaka reza.

Hipoteze su potvrđene

Na osnovu pothipoteza i na osnovu urađene disertacije može se zaključiti:

I Potvrđeno je da raspodele podataka utiču na algoritam diskretizacije.
II Potvrđeno je da raspodele podataka utiču na kvalitet klisifikacije.
III Potvrđeno je da diskretizacija podataka može značajno da utiče na inkonzistentnost tabele kod podataka sa određenim raspodelama.
IV Potvrđeno je da je na osnovu raspodele podataka moguće smanjiti broj tačaka diskretizacije tako da redukt skupa tabelarno organizovanih podataka ostane nepromenjen.

Potvrđena je glavna hipoteza ukupnim rezultatima ove disertacije:

Moguće je izgraditi algoritam za aproksimativne diskretizacije
u odnosu na raspodele iz tabelarno organizovanih podataka.

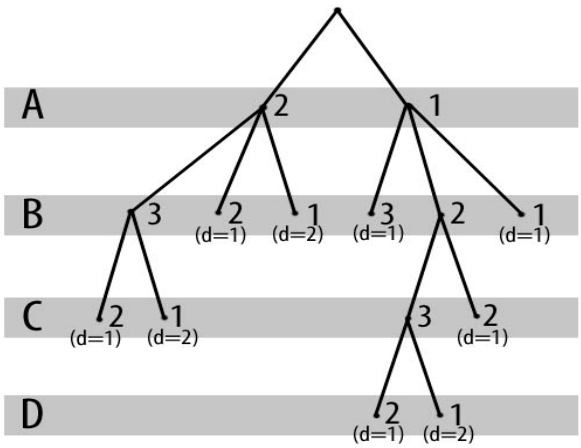
Kao bitan rezultat ove disertacije predviđena je modifikacija bar jednog algoritma za diskretizaciju podataka, tako da omogući rad sa aproksimativnim vrednostima na osnovu raspodela podataka. Konstrukcija algoritma aproksimativne diskretizacije APPROX MD u odnosu na rezultate diskretizacije algoritma maksimalne razberivosti,

u odnosu na konzistentnost diskretizovane tabele i u odnosu na predloženu redukciju tačaka reza predstavlja traženi rezultat.

8.2 SMERNICE BUDUĆIH ISTRAŽIVANJA

Radom na ovoj disertaciji sagledana je mogućnost implementacije algoritma APPROX MD u okviru sistema SSCO koji je razvijen na Tehničkom fakultetu „Mihajlo Pupin“ u Zrenjaninu i koji je korišćen i razvijan u nekoliko projekata [Brtka, 2008]. Za implementaciju algoritma APPROX MD potrebno je definisati odgovarajuće matrice susedstva koje bi bile „kompatibilne“ teoriji SSCO sistema i matrici razberivosti koja se koristi u okviru algoritma maksimalne razberivosti MD. Zbog različitih načina grafovskog predstavljanja određene relacije, potrebno je uraditi istraživanje predstavljanja kako relacije razberivosti tako i drugih relacija (bitnih za rad u teoriji grubih skupova) u okviru sistema SSCO.

Sistem SSCO je povezan i sa stablima odlučivanja i sa teorijom grubih skupova a baziran je na specifičnim iteracijama po grafu koji je konceptualno drugačiji od grafova koji se generišu kako u stablima odlučivanja tako i u teoriji grubih skupova (slika 60).



Slika 60.

Pretpostavlja se da bi na osnovu određene konstrukcije matrice susedstva grafa u SSCO sistemu bilo moguće povezati je sa podacima histograma. To bi možda dodatno uticalo na modifikaciju algoritma FixedPoins.

Implementacija algoritma diskretizacije APPROX MD je moguća i u sistemima klasifikacije baziranim na teoriji grubih skupova koji koriste algoritam maksimalne razberivosti za diskretizaciju podataka.

LITERATURA

Banknote Authentication (2011), https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

Bay S. D. (2015), Multivariate Discretization of Continuous Variables for Set Mining, Department of Information and Computer Science, University of California, Irvine, http://www.ime.unicamp.br/~wanderson/Artigos/multivariate_discretization_of_continuous_variables.pdf.

Bazan J. G., Nguyen H. S., Nguyen S. H., Synak P., Wroblewski J. (2000), Rough Set Algorithms in Classiffcation Problem, Rough Set Methods and Applications, ISBN 978-3-662-00376-3, Volume 56 of the series Studies in Fuzziness and Soft Computing pp 49-88.

Bazan J., Nguyen H. S., Skowron A., and Szczuka M. (2003), A View on Rough Set Concept Approximations, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Proceedings of RSFDGrC 2003, volume 2639 of Lecture Notes in Artificial Intelligence, pages 181–188.

Bazan J.G., Nguyen H.S., Nguyen S.H., Synak P., Wróblewski J. (2000), Rough Set Algorithms in Classification Problem, Rough Set Methods and Applications, Volume 56 of the series Studies in Fuzziness and Soft Computing, ISBN 978-3-662-00376-3, pp 49-88.

Birge, L., i Rozenholc, Y. (2002). How many bins should be put in a regular histogram, Prepublication no 721, Laboratoire de Probabilites et Modeles Aleatoires, CNRS-UMR 7599, Universite Paris VI & VII.

Blood (2008), Blood Transfusion Service Center Data Set, http://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

Branović Ž. (2003), Uvod u teoriju informacija i komunikacija, Univerzitet u Novom Sadu, Tehnički fakultet "Mihajlo Pupin", Zrenjanin, ISBN: 86-80711-31-4.

Breast Cancer Wisconsin (Original) Data Set (2011), http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29.

Brtka E., Ognjenovic V, Brtka V. (2012), The evaluation of the overall knowledge of the students by usage Dynamic Reducts, Technics Technologies Education Management, Vol7No4, ISSN: 1840-1503, pp. 1672-1680.

Brtka V. (2008), Automatska sinteza baze pravila u inferentnim sistemima, doktorska disertacija, Univerzitet u Novom Sadu, Tehnički fakultet „Mihajlo Pupin" Zrenjanin.

Brtka V., Brtka E., Ognjenovic V. and Berkovic I. (2011a), The Decision Rules Synthesis Based on Similarity Relation, SCIENTIFIC BULLETIN of The "POLITEHNICA" University of Timişoara, Romania, Transactions on AUTOMATIC CONTROL and COMPUTER SCIENCE, Vol. 56 (70), No. 3, ISSN 1224-600X, pp. 97-104.

Brtka V, Ognjenovic V., Brtka E., Berkovic I. (2011b), The Rough Sets Based Data Analysis in Small and Medium Sized Enterprises, 6th IEEE International

Symposium on Applied Computational Intelligence and Informatics - SACI 2011, Timisoara, Romania, pp. 373-378.

Brtka V., Stokic E., Srdic B. (2008), Automated extraction of decision rules for leptin dynamics—A rough sets approach, Journal of Biomedical Informatics 41, pp. 667 – 674.

C4.5 Tutorial, (2016), Machine Learning/Decision Trees/

Cameron C. (2015), EXCEL 2007: Histogram, Dept. of Economics, Univ. of Calif., http://cameron.econ.ucdavis.edu/excel/ex11histogram.html

Cardiotocography (2015), http://archive.ics.uci.edu/ml/datasets/Cardiotocography

Catlett, J. (1991), On changing continuous attributes into ordered discrete attributes, in Y. Kodratoff, ed., Proceedings of the European Working Session on Learning, Berlin, Germany: Springer-Verlag, pp. 164-178.

Challenger (1993), Challenger USA Space Shuttle O-Ring Data Set http://archive.ics.uci.edu/ml/datasets/Challenger+USA+Space+Shuttle+O-Ring

Chang J.H., Fan Kuo.C., Chang Y.L. (2002), Multi-modal gray-level histogram modeling and decomposition, Image and Vision Computing 20, pp. 203-216.

Chaudhuri D., Agrawal A. (2010), Split-and-merge Procedure for Image Segmentation using Bimodality Detection Approach, Defence Science Journal, 2010, 60(3), pp.290-301.

Chebrolu S., Sanjeevi S.G. (2015), Attribute Reduction on Continuous Data in Rough Set Theory using Ant Colony Optimization Metaheuristic, WCI '15 Proceedings of the Third International Symposium on Women in Computing and Informatics, ISBN: 978-1-4503-3361-0, Pages 17-24.

Cong R., Wang X., Li K., and Yang N. (2010), New method for discretization of continuous attributes in rough set theory, Journal of Systems Engineering and Electronics Vol. 21, No. 2, pp.250–253

Delon J., Desolneux A., Lisani J-L. and Petro A-B. (2007), A non parametric approach for histogram segmentation, IEEE Transactions on Image Processing, Vol. 16, No. 1, pp. 253-261.

Delon J., Desolneux A., Lisani J-L. et Petro A-B. (2005 a), Automatic Color Palette, ICIP 2005, Genova.

Delon J., Desolneux A., Lisani J-L. et Petro A-B. (2005 b), Color Image Segmentation using Acceptable Histogram Segmentation, IbPRIA05 proceedings, part II, Springer LNCS series, pp.239-246.

Delon J., Desolneux ALisani., J-L. and Petro A-B. (2007), Automatic Color Palette, Inverse Problems and Imaging(IPI), vol.1, no.2, pp.265-287.

Dougherty J., Kohavi R., and Sahami M. (1995). Supervised and unsupervised discretization of continuous features. In A. Prieditis and S. Russell, editors, Proc. Twelfth International Conference on Machine Learning, pages 194–202. Morgan Kaufmann.

Du W. S., Hu B. Q. (2015), Allocation Reductions in Inconsistent Decision Tables Based on Dominance Relations, Fuzzy Information and Engineering, Volume 7, Issue 3, September 2015, Pages 259–273.

Düntsch I., Gedigab G. (1998), Uncertainty measures of rough set prediction, Artificial Intelligence Volume 106, Issue 1, Pages 109–137.

EasyFit software (2015), EasyFit software :: Product Specification, http://www.mathwave.com/products/easyfit_desc.html

Fayyad U.M. i Irani K.B. (1992), On the Handling of Continuous-Valued Attributes in Decision Tree Generation, Machine Learning, v.8, pp.87-102.

Fayyad U.M., i Irani K.B. (1993), The Attribute Selection Problem in Decision Tree Generation, Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022-1027.

Ferreira A. J., Figueiredo M.A.T. (2014), Incremental filter and wrapper approaches for feature discretization, Neurocomputing, Volume 123, Pages 60-74.

Gama J, Pinto C. (2006), Discretization from Data Streams: Applications to Histograms and Data Mining, SAC '06 Proceedings of the 2006 ACM symposium on Applied computing, Pages 662-667.

Gama J., Torgo L., Soares C. (2015), Dynamic Discretization of Continuons Attributes, www.liaad.up.pt/~ltorgo/Papers/DDCA.ps.gz.

García S., Luengo J., Sáez J.A., López V. and Herrera F. (2013), A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning, IEEE Transactions on Knowledge & Data Engineering vol.25 Issue No.04.

Glass Identification (2014), http://archive.ics.uci.edu/ml/datasets/Glass+Identification.

Grzymala-Busse J.W. (2005), Rough Set Theory with Applications to Data Mining, Real World Applications of Computational Intelligence, Volume 179 of the series Studies in Fuzziness and Soft Computing pp 221-244.

Guillemot T., Delon J. (2016), Midway Image Equalization, preprint du journal Image Processing On Line.

Haberman (1999), Haberman's Survival Data Set http://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

Han J., Kamber M., Pei J. (2011), Data Mining: Concepts and Techniques, Elsevier, Third Edition, The Morgan Kaufmann Series in Data Management Systems, ISBN-13: 978-0123814791.

Iris (1936), Iris Data Set, https://archive.ics.uci.edu/ml/datasets/Iris

Ismail M. K., Ciesielski V. (2003), An Empirical Investigation of the Impact of. Discretization on Common Data. Distributions., Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03): Design and Application of Hybrid Intelligent Systems.

Jiang F., Zhao Z., Ge Y. (2010), A Supervised and Multivariate Discretization Algorithm for Rough Sets, Lecture Notes in Computer Science, Volume 6401/2010, 596-603.

Jiang S.Y., Li X., Zheng Q., Wang L. X. (2009), Approximate Equal Frequency Discretization Method, 2009 WRI Global Congress on Intelligent Systems (Volume:3), ISBN: 978-0-7695-3571-5, pp.514 – 518.

Johnson D. S. (1974). Approximation algorithms for combinatorial problems. Journal of Computer and System Sciences, 9:256–278.

Komorowski J., Pawlak Z., Polkowski L., Skowron A. (1999), Rough Sets: A Tutorial, in S.K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization. A New Trend in Decision-Making, Springer-Verlag, Singapore, 3-98.

Kontkanen P., Myllymaki P. (2007), MDL Histogram Density Estimation, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics March 21-24, 2007, San Juan, Puerto Rico, 2:219-226.

Liu H, Liu D.Y., Shi X.H., Gao Y. (2008), An attribute discretization algorithm based on Rough Set and information entropy, 2008 International Conference on Machine Learning and Cybernetics, ISBN 978-1-4244-2095-7.

Liu Q., Chen L., Zhang J., and Min F. (2006), Knowledge Reduction in Inconsistent Decision Tables, Advanced Data Mining and Applications, Volume 4093 of the series Lecture Notes in Computer Science pp 626-635.

Lover R. (2008), Elementary Logic: For Software Development, Springer Science & Business Media, ISBN: 978-1-84800-081-0

Matlab - Histogram (2015), Histogram with a distribution fit, http://www.mathworks.com/help/stats/histfit.html

Milikić N. (2016), Klasifikacija – Stabla odlučivanja, Laboratorija za veštačku inteligenciju, FON, Beograd, http://nikola.milikic.info.

Muhlenbach F., Rakotomalala R. (2005), "Discretization of Continuous Attributes", in J. Wang, editor, Encyclopedia of Data Warehousing and Mining, Idea Group Reference, pp.397-402.

Nguyen H. S. and Slezak D. (1999), Approximate Reducts and Association Rules – Correspondence and Complexity Results, New Directions in Rough Sets, Data Mining and Granular-Soft Computing (Proc. of RSFDGrC'99, Yamaguchi, Japan), volume 1711 of LNAI 1711, pages 137–145. Springer, Heidelberg, Germany.

Nguyen H.S. (1997), Discretization of Real Value Attributes, Boolean Reasoning Approach. PhD thesis, Warsaw University, Warsaw, Poland.

Nguyen H.S. (1998), Discretization problem for rough sets methods. In: Polkowski, L., Skowron, A. (eds.) RSCTC 1998. LNCS (LNAI), vol. 1424, pp. 545–555. Springer, Heidelberg.

Nguyen H.S. (2006 b), Approximate Boolean Reasoning: Foundations and Applications in Data Mining, Institute of Mathematics, Warsaw University, Institute of Mathematics, Warsaw University.

Nguyen H.S. (2006), Approximate boolean reasoning: foundations and applications in data mining, Transactions on rough sets V, 334-506.

Nguyen H.S., Nguyen S.H. (1998), Discretization Methods in Data Mining. In: Rough Sets in Knowledge Discovery, Physica, pp. 451–482.

Nguyen H.S., Skowron A. (1997), Boolean reasoning for feature extraction problems. In Z. W. Ra´s, A. Skowron, editors, Proceedings of 10th International Symposium on the Foundations of Intelligent Systems (ISMIS'97), LNAI 1325, 117–126, Springer, Heidelberg.

Nguyen H.V., Müller E., Vreeken J., Böhm K. (2014), Unsupervised interaction-preserving discretization of multivariate data, Data Mining and Knowledge Discovery.

Nguyen S. H. and Nguyen H. S. (1996), Some Efficient Algorithms for Rough Set Methods. In Proceedings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'96, pages 1451–1456, Granada, Spain.

Ognjenović V., Brtka V., Berković I. (2013), Discretization influence on data reduction, International Conference on Applied Internet and Information Technologies ICAIIT 2013, Zrenjanin, Proceedings, ISBN 978-86-7672-211-2, pp. 158-161.

Ognjenovic V., Brtka V., Berkovic I., Brtka E. (2012), Comparison of the classification rules generated by See 5.0 and SSCO Systems, Proceedings of the 23nd Central European Conference on Information and Intelligent Systems - CECIIS 2012, Varaždin, Croatia, ISSN 1847-2001, pp. 71-76.

Ognjenović V., Brtka V., Brtka E., Berković I. (2016), Diskretizacija podataka redukcijom tačaka reza, INFOTEH-JAHORINA Vol. 15, pp. 665-670.

Ognjenovic V., Brtka V., Jovanovic M., Berkovic I. (2011 a), Supervised Discretization for Rough Sets – a Neighborhood Graph Approach, , Proceedings of the 22nd Central European Conference on Information and Intelligent Systems - CECIIS 2011, Varaždin, Croatia, ISSN 1847-2001, pp. 265 – 272.

Ognjenovic V., Brtka V., Jovanovic M., Brtka E., Berkovic I. (2011 b), The Representation of Indiscernibility Relation by Graph, Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium, Subotica, Serbia, IEEE Catalog Number: CFP1184C-CDR, ISBN: 978-1-4577-1973-8, pp. 91-94.

Øhrn A. (2001), ROSETTA Technical Reference Manual, Department of Computer and Information Science, Norwegian University of Science and Technology, www.lcb.uu.se/tools/rosetta/materials/manual.pdf

Øhrn A., Komorowski J., Skowron A., and Synak P. (1998), The ROSETTA, software system. In L. Polkowski and A. Skowron, editors, Rough Sets in Knowledge Discovery 2. Applications, Case Studies and Software Systems, number 19 in Studies in Fuzziness and Soft Computing, pages 572–576.Physica-Verlag, Heidelberg, Germany.

Patwardhan S.C. (2016), Polynomial Approximation, http://nptel.ac.in/courses/103101009/14.

Pawlak Z. (1982), Rough sets, International Journal of Computer and Information Sciences, 11, 341-356.

Pawlak Z. (2002), Rough set theory and its applications, Journal of Telecommunications and Information Technology, 3 2002, vol. 3, pp. 7-10.

Pawlak, Z. (1991), Rough sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

Pawlak, Z. (2001), Rough Sets and their Applications, Computational Intelligence in Theory and Practice, ISBN 978-3-7908-1831-4, pp. 73-91.

Polkowski L. (2003), Rough Sets and Current Trends in Computing: First International Conference, RSCTC'98 Warsaw, Poland, June 22–26, 1998 Proceedings, Springer, May 20, 2003.

Qian Y., Liang J., Li D., Wang F, Ma N. (2010 a), Approximation reduction in inconsistent incomplete decision tables, Knowledge-Based Systems, Volume 23, Issue 5, July 2010, Pages 427-433, ISSN 0950-7051, http://dx.doi.org/10.1016/j.knosys.2010.02.004.

Qian Y., Liang J., Pedrycz W., Dang C. (2010 b), Positive approximation: An accelerator for attribute reduction in rough set theory, Artificial Intelligence, Volume 174, Issues 9–10, Pages 597-618.

Quinlan R. (2014), C4.5 Example, http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/c4.5_prob1.html.

Ramırez-Gallego S., Garcıa S., Mourino-Talın H., Martınez-Rego D., Bolon-Canedo V., Alonso-Betanzos A., Benıtez J. M., Herrera F. (2016), Data Discretization: Taxonomy and Big Data Challenge, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 6, Issue 1, pages 5–21.

RSES (2014), RSES 2.2 User's Guide, Warsaw University, http://logic.mimuw.edu.pl/~rses/RSES_doc_eng.pdf.

Runić J.M. (2016), INFORMACIONI SISTEMI ZA PODRŠKU MENADŽMENTU, http://odlucivanje.fon.bg.ac.rs/wp-content/uploads/Classification-Tree-K-NN-Test-learners-Predictions.pdf

Schmidberger, G. i Frank, E. (2005). Unsupervised discretization using tree-based density estimation. In A. Jorge et al. (Eds), Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. (pp. 240-251). Berlin: Springer

Sengupta S., Das A. K. (2014), A STUDY ON ROUGH SET THEORY BASED DYNAMIC REDUCT FOR CLASSIFICATION SYSTEM OPTIMIZATION, International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 5, No. 4.

Shannon, C. E. (1951), Prediction and Entropy of Printed English,. Bell System Technical Journal 30 (1): 50–64.

Sheng-yi J. , Xia L. ; Qi Z ; Lian-xi W. (2009), Approximate Equal Frequency Discretization Method, Sch. of Inf. Guangdong, Univ. of Foreign Studies, Guangzhou, China, 2009 WRI Global Congress on Intelligent Systems

Shi Z., Xia Y., Wu F. and Dai J.(2014), The Discretization Algorithm for Rough Data and Its Application to Intrusion Detection, JOURNAL OF NETWORKS, VOL. 9, NO. 6.

Skowron A. and Rauszer C. (1992). The discernibility matrices and functions in information systems, Intelligent Decision Support, Volume 11 of the series Theory and Decision Library, ISBN 978-90-481-4194-4, pp 331-362.

Skowron A., Nguyen H.S. (1999), Boolean reasoning scheme with some applications in data mining, Principles of Data Mining and Knowledge Discovery, 107-115.

Slezak D. (2000), Various Approaches to Reasoning with Frequency Based Decision Reducts: A Survey, Rough Set Methods and Applications, Volume 56 of the series Studies in Fuzziness and Soft Computing pp 235-285.

Slezak D. (2002), Approximate Entropy Reducts, Fundamenta Informaticae, Volume 53 Issue 3-4, Pages 365-390.

Statlog (2015), (Australian Credit Approval) https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29

Suraj Z. (2004), An Introduction to Rough Set Theory and Its Applications, A tutorial, ICENCO'2004, December 27-30, 2004, Cairo, Egypt, http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=B1351C4BE7AA7AE958940178DE88958F?doi=10.1.1.488.192&rep=rep1&type=pdf.

Tang X., Shu L. (2014), ATTRIBUTE REDUCTION ALGORITHM BASED ON COGNITIVE MODEL OF GRANULAR COMPUTING IN INCONSISTENT DECISION INFORMATION SYSTEMS, Technical Gazette, Vol.21 No.1 February 2014.

UCI (2015), UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.html

Virginia G. (2013), Lexicon-based Document Representation, Fundamenta Informaticae 124, pp. 27–46.

Webb G.I. (2014), Contrary to Popular Belief Incremental Discretization can be Sound, Computationally Efficient and Extremely Useful for Streaming Data, , 2014 IEEE International Conference on Data Mining, ISBN 978-1-4799-4303-6, Pages 1031 – 1036.

Wei J.M. (2003), ROUGH SET BASED APPROACH TO SELECTION OF NODE, International Journal of Computational Cognition, Volume 1, Number 2, Pages 25–40.

Wilt Data Set (2015), http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice.

Zhang Y.; Cheung Y.M. (2014), Discretizing Numerical Attributes in Decision Tree for Big Data Analysis,. Data Mining Workshop (ICDMW), 2014 IEEE International Conference.

Zhao J., Zhou Y. (2009), New heuristic method for data discretization based on rough set theory, The Journal of China Universities of Posts and Telecommunications, Volume 16, Issue 6, Pages 113-120.

DODACI

1. DODACI ZA ANALIZU BAZE IRIS

Baza Iris je vrlo korišćena i vrlo poznata baza u okviru literature prepoznavanja paterna (*pattern recognition*). Kreirao ju je R.A. Fisher 1936. godine, a donirao ju je Michael Marshall [Iris, 1936]. Baza sadrži 3 klase od kojih je svaka od 50 instanci, i svaka klasa se odnosi na tip biljke iris.

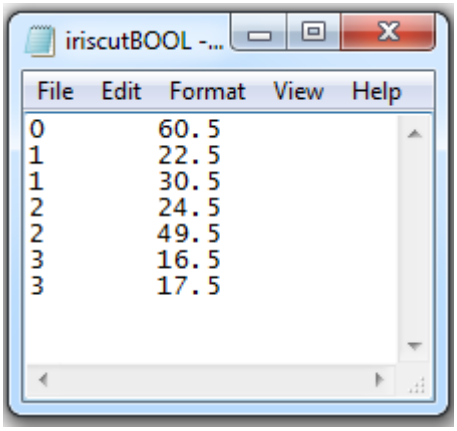
Informacije o atributima (preuzeto iz [Iris, 1936]):

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Broj instanci je 150. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

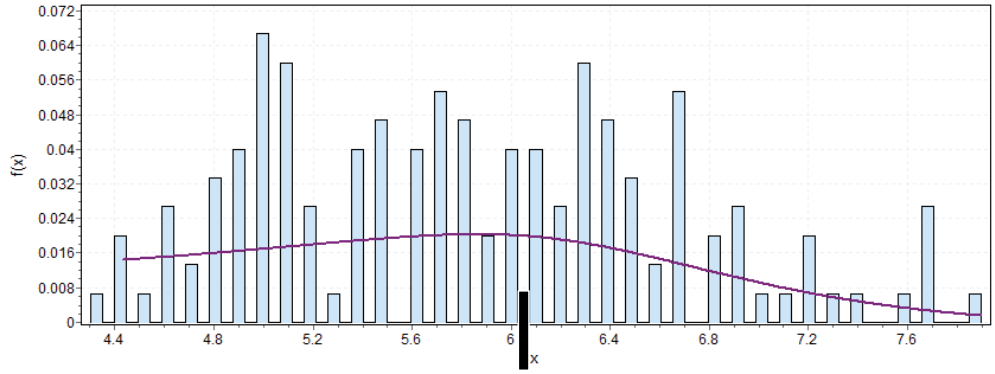
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Iris algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 61. Za prvi atribut kojem je sistem Rosetta dodelio oznaku nula, postoji samo jedna tačka reza sa vrednošću 60.5; za drugi atribut sa oznakom jedan postoje dve tačke reza sa vrednostima 22.5 i 30.5 i tako redom.

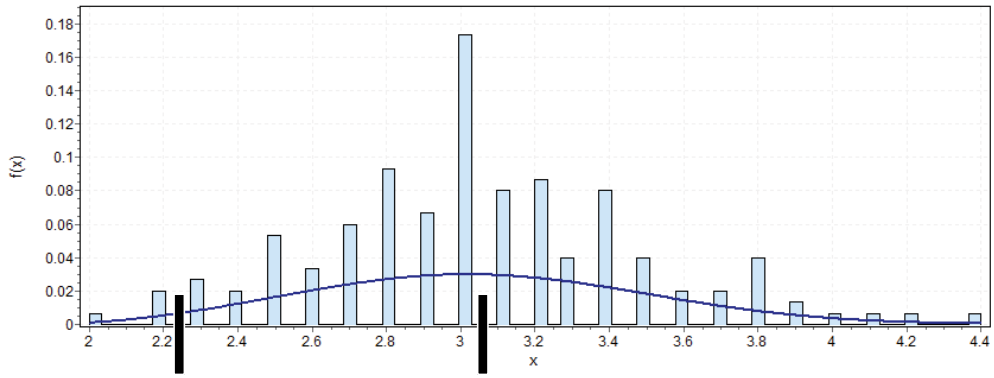


Slika 61. Tačke reza baze Iris dobijene algoritmom maksimalne razberivosti

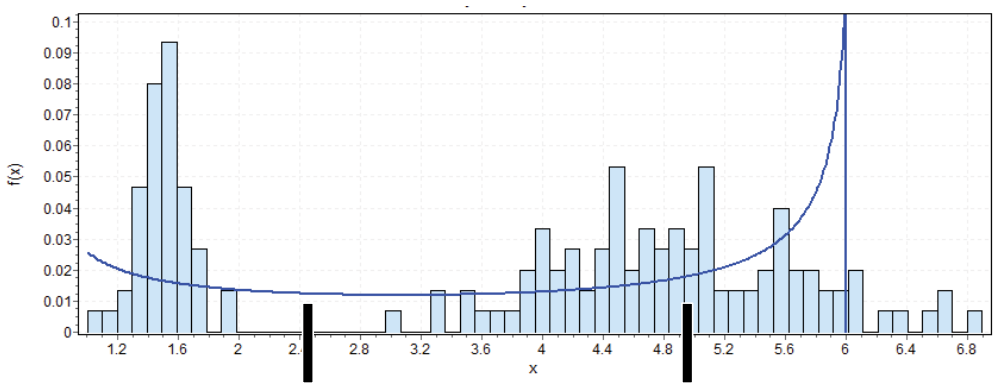
U odnosu na raspodelu podataka baze Iris koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 62 - 65).



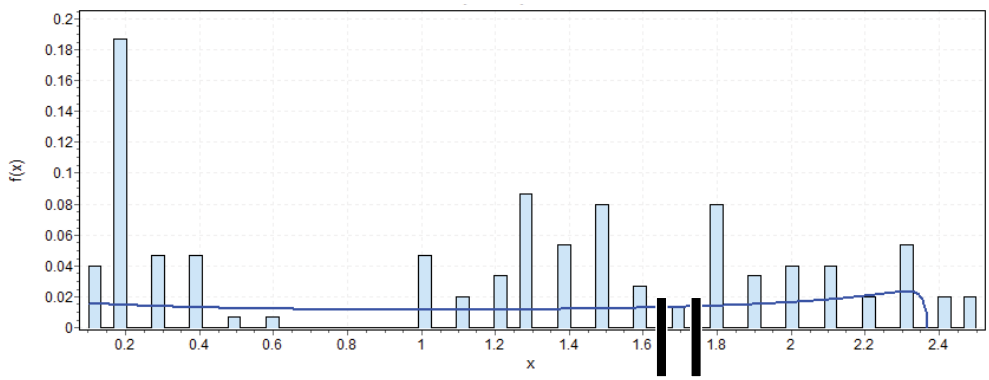
Slika 62. Raspodela podataka 1. atributa sepal length (baze Iris) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



Slika 63. Raspodela podataka 2. atributa sepal width (baze Iris) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



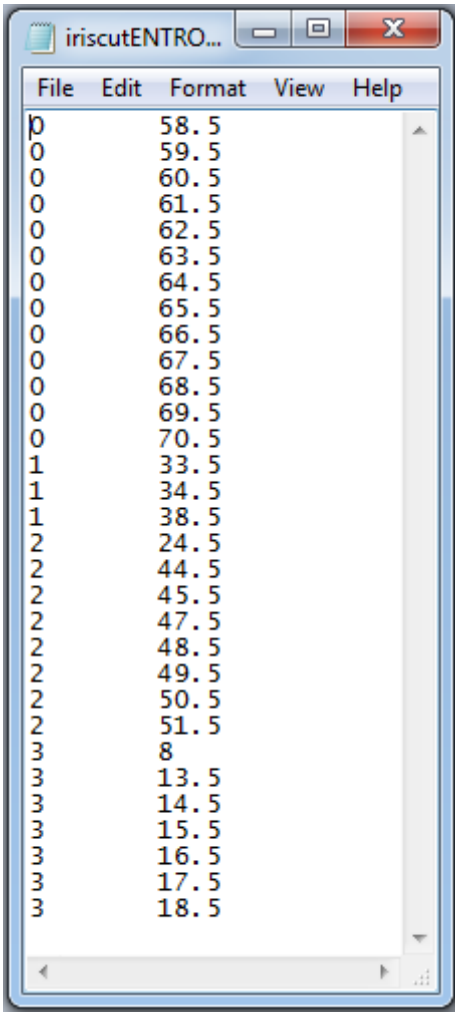
Slika 64. Raspodela podataka 3. atributa petal lenght (baze Iris) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 65. Raspodela podataka 4. atributa petal width (baze Iris) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

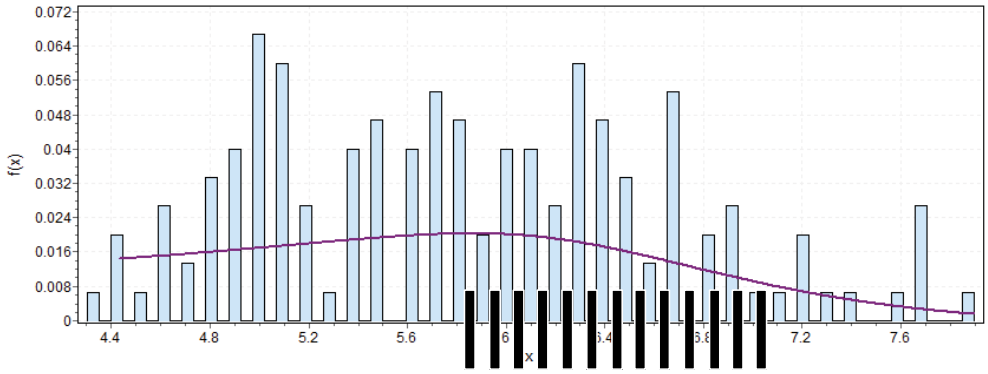
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Iris algoritmom baziranim na entropiji, dobijene su tačke reza koje su prikazane na slici 66. Za prvi atribut kojem je sistem Rosetta dodelio oznaku nula, postoji 13 reza, za drugi atribut čija je oznaka jedan postoje 3 tačke reza i tako redom

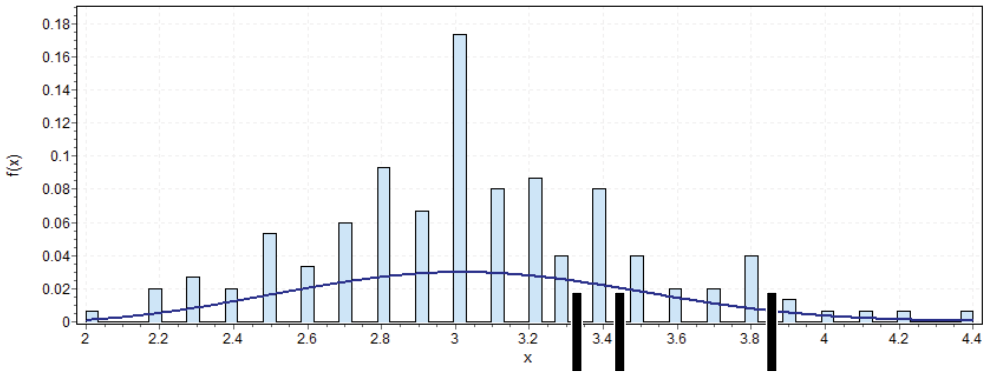


Slika 66. Tačke reza baze Iris dobijene algoritmom baziranim na entropiji

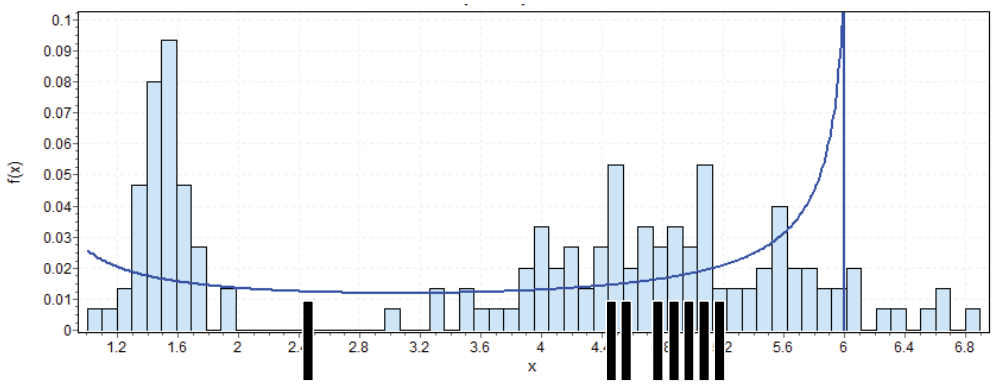
Broj tačaka reza je značajno veći kod tri atributa u odnosu na rezultate dobijene algoritmom maksimalne razberivosti. Kod drugog atributa koji ima oznaku jedan, dobijene su tri tačke reza. U odnosu na raspodelu podataka baze Iris koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 67 - 70).



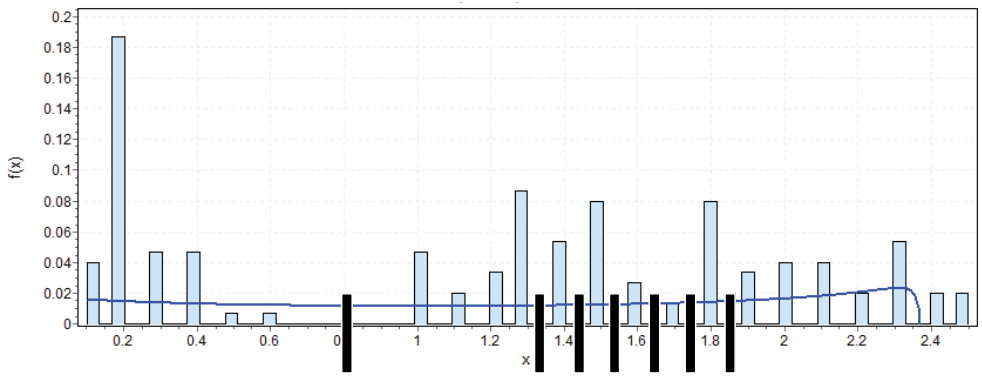
Slika 67. Raspodela podataka 1. atributa sepal length (baze Iris) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 68. Raspodela podataka 2. atributa sepal width (baze Iris) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 69. Raspodela podataka 3. atributa petal lenght (baze Iris) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 70. Raspodela podataka 4. atributa petal width (baze Iris) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji

2. DODACI ZA ANALIZU BAZE BLOOD TRANSFUSION SERVICE CENTER

Baza Blood Transfusion Service Center je dobijena iz Hsin-Chu grada sa Tajvana. Vlasnik i donor je Prof. I-Cheng Yeh sa Department of Information Management, Chung-Hua University, Tajvan. Baza je sačinjena od podataka slučajno izabranih 748 donora krvi, a donirana je 2008. godine [Blood, 2008].

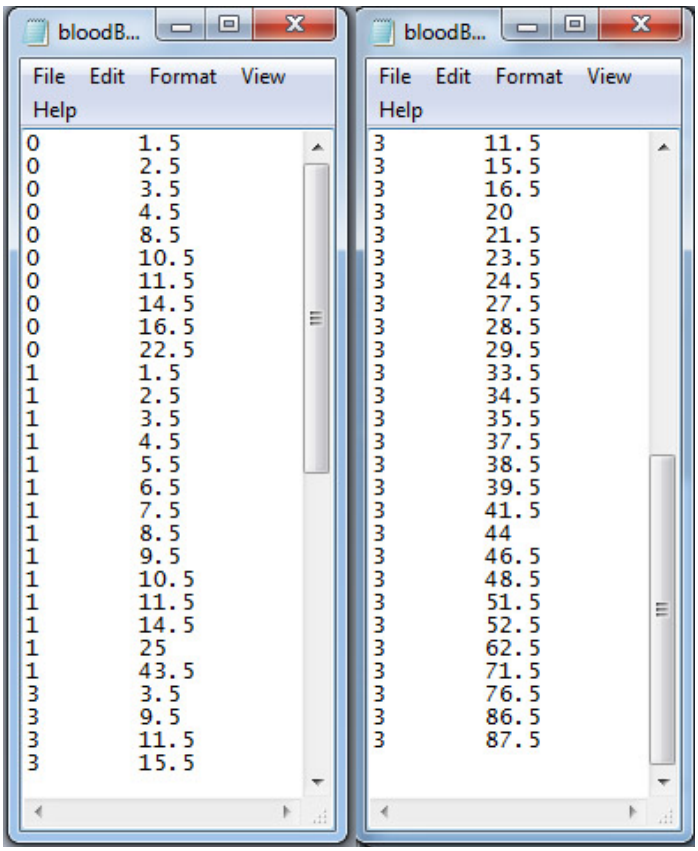
Informacije o atributima (preuzeto iz [Blood, 2008]):

1. R (Recency - months since last donation),
2. F (Frequency - total number of donation),
3. M (Monetary - total blood donated in c.c.),
4. T (Time - months since first donation), and
5. a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

Broj instanci je 748. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

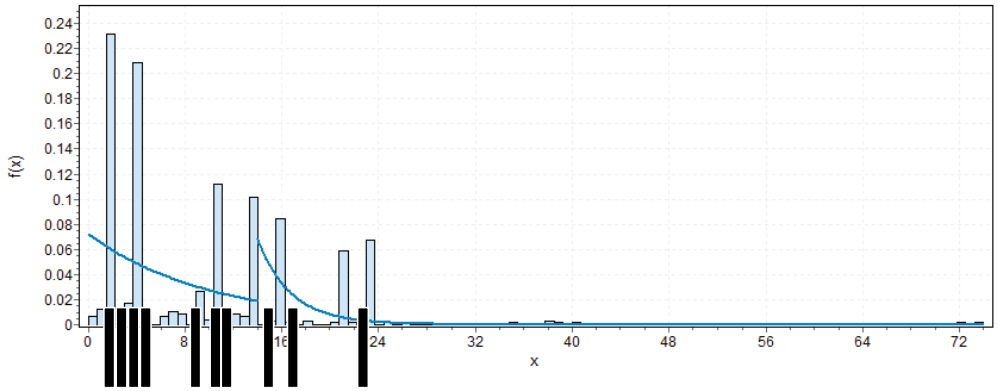
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Blood Transfusion Service Center algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 71. Algoritam nije dao ni jednu tačku reza drugog atributa.

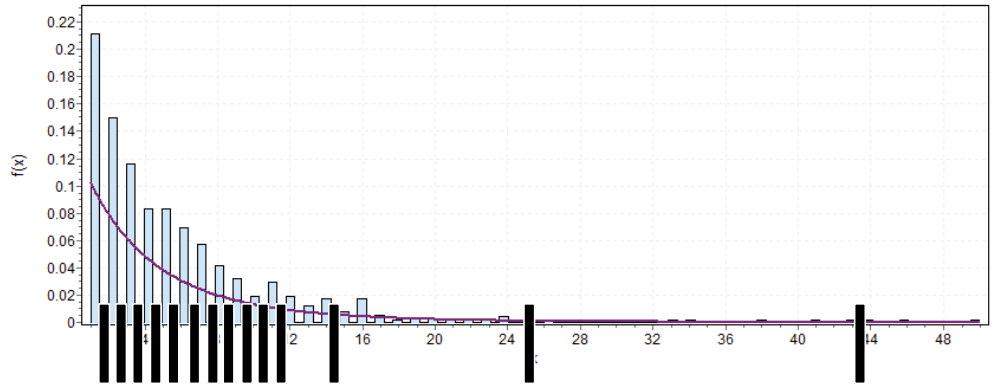


Slika 71. Tačke reza baze Blood Transfusion Service Center dobijene algoritmom maksimalne razberivosti

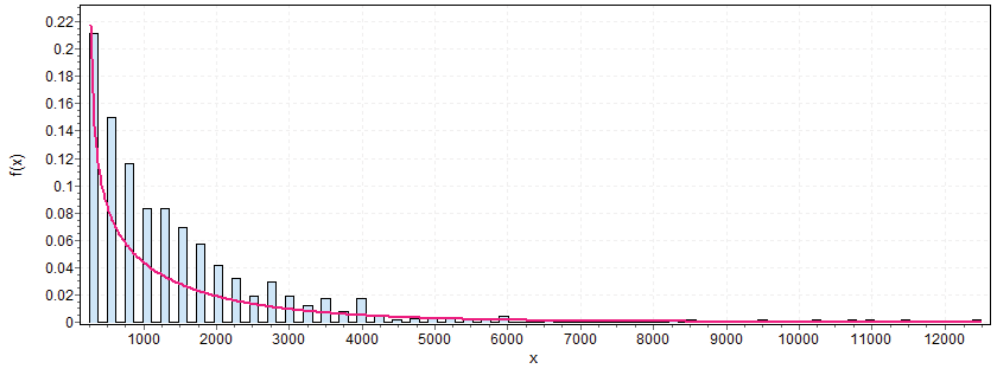
U odnosu na raspodelu podataka baze Blood Transfusion Service Center koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 72 - 75).



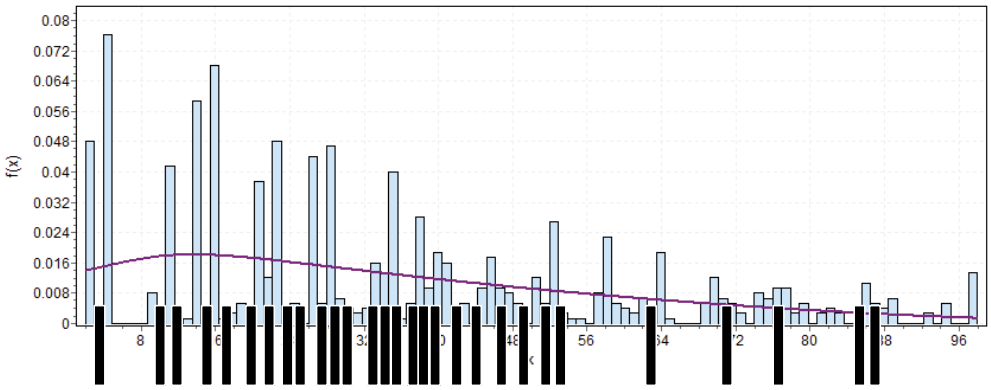
Slika 72. Raspodela podataka 1. atributa R (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 73. Raspodela podataka 2. atributa F (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



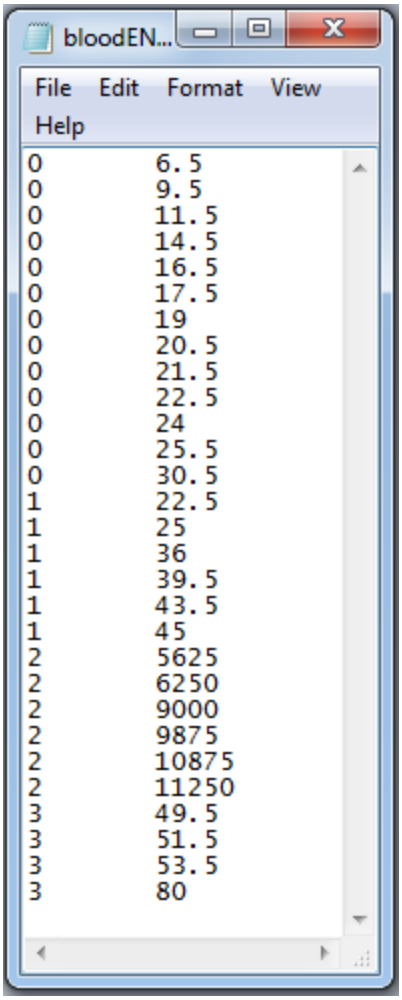
Slika 74. Raspodela podataka 3. atributa M (baze Blood Transfusion Service Center) bez tačke reza na osnovu algoritma maksimalne razberivosti



Slika 75. Raspodela podataka 4. atributa T (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

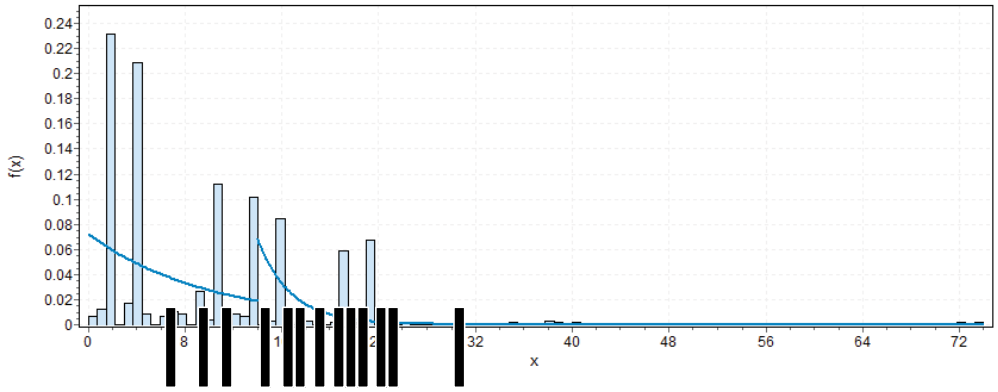
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Blood Transfusion Service Center algoritmom baziranim na entropiji, dobijene su tačke reza koje su prikazane na slici 76. Za prvi atribut kojem je sistem Rosetta dodelio oznaku nula, tačke reza su 6.5, 9.5, 11.5, ..., 30.5. Za drugi atribut čija je oznaka jedan tačke reza su 22.5, 25, ..., 45, i tako redom.

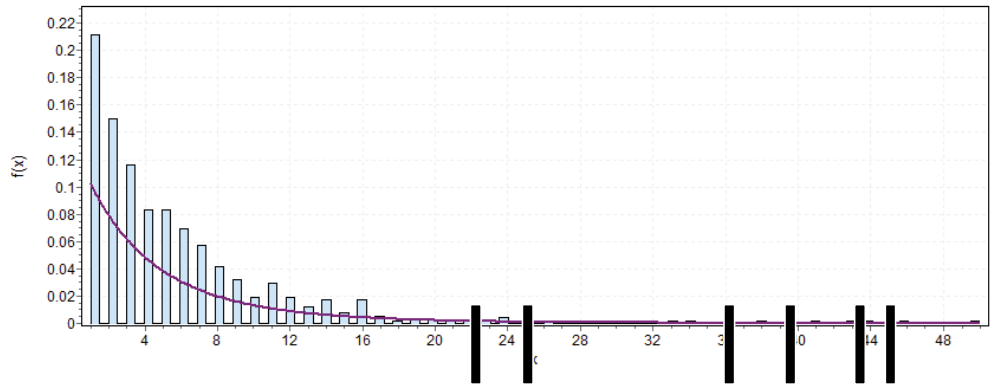


Slika 76. Tačke reza baze Blood Transfusion Service Center dobijene algoritmom baziranim na entropiji

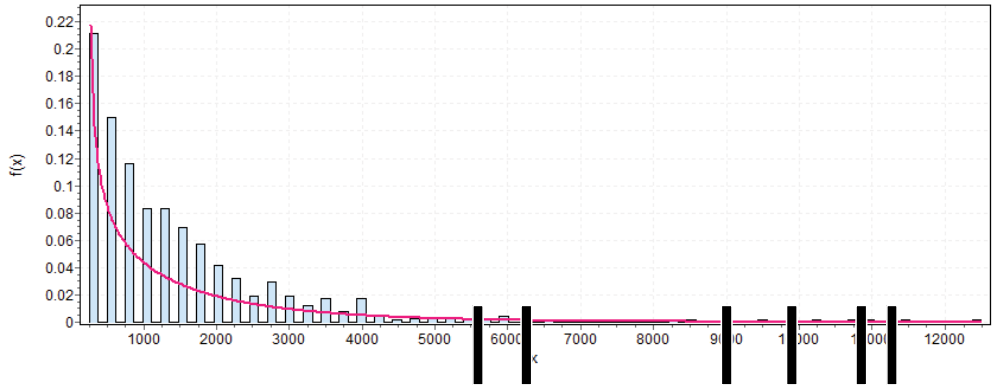
U odnosu na tačke reza dobijene algoritmom maksimalne razberivosti, u ovom slučaju, primenom algoritma baziranog na entropiji, postoje tačke reza trećeg atributa. U odnosu na raspodelu podataka baze Blood Transfusion Service Center koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 77 - 80).



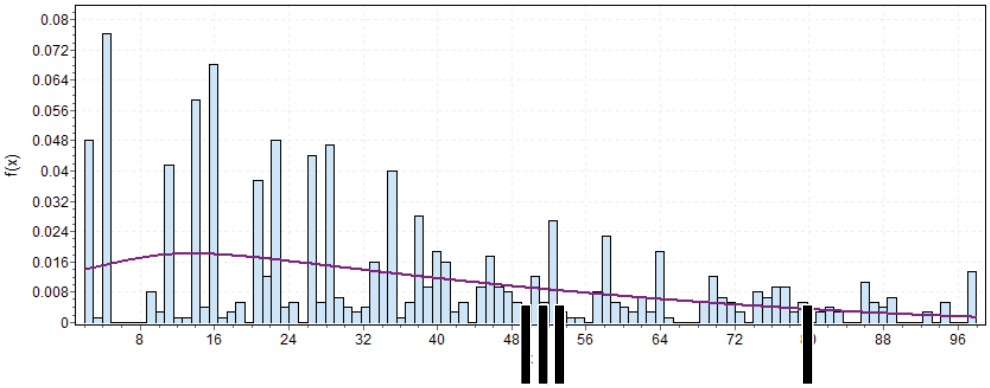
Slika 77. Raspodela podataka 1. atributa R (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 78. Raspodela podataka 2. atributa F (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 79. Raspodela podataka 3. atributa M (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 80. Raspodela podataka 4. atributa T (baze Blood Transfusion Service Center) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji

3. DODACI ZA ANALIZU BAZE BANKNOTE AUTHENTICATION

Baza Banknote Authentication je dobijena sa University of Applied Sciences, Ostwestfalen-Lippe, Nemačka, 2012. godine. Vlasnik baze je Volker Lohweg, University of Applied Sciences, Ostwestfalen-Lippe, a donor je Helene Darksen, University of Applied Sciences, Ostwestfalen-Lippe. Podaci ove baze su preuzeti sa slika koje su uzete sa originalnih i falsifikovanih novčanica. Za digitalizaciju je korišćena industrijska kameta za pregled štampe. Slike imaju 400 x 400 piksela, a zbog objektiva i udaljenosti koja bi istraživala objekat u sivim tonovima, ostvarena je rezolucija od 660 dpi. Za izdvajanje osobina sa slike korišćene su Wavelet transformacije.

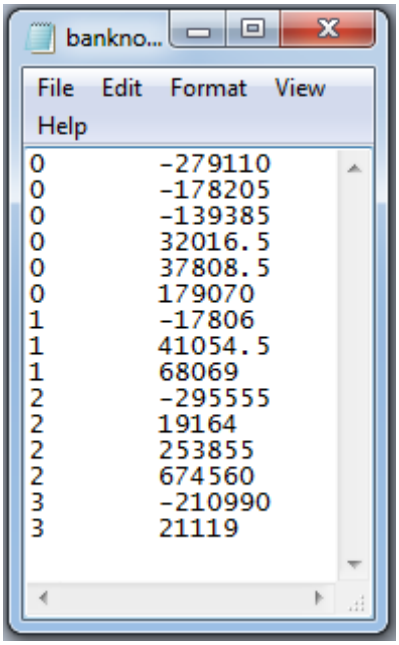
Informacije o atributima (preuzeto iz [Banknote, 2012]):

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer).

Broj instanci je 1372. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

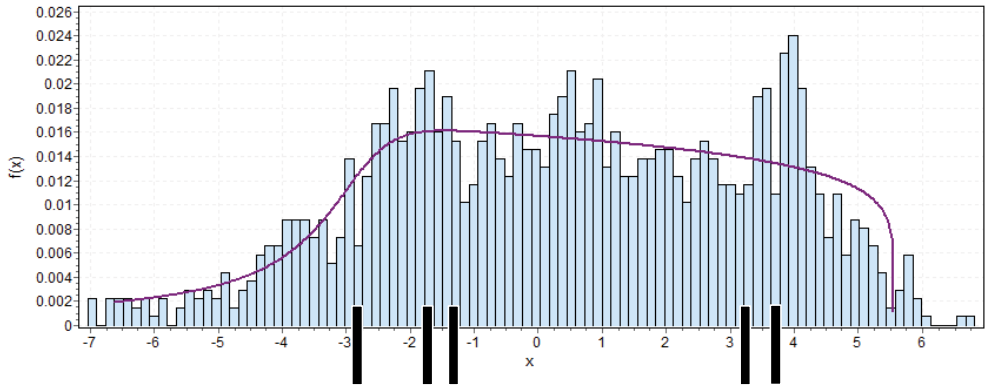
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Banknote Authentication algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 81.

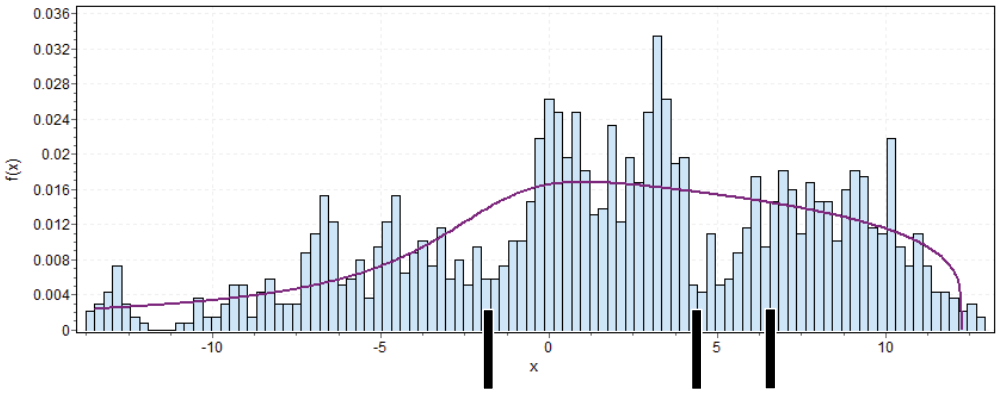


Slika 81. Tačke reza baze Banknote Authentication dobijene algoritmom maksimalne razberivosti

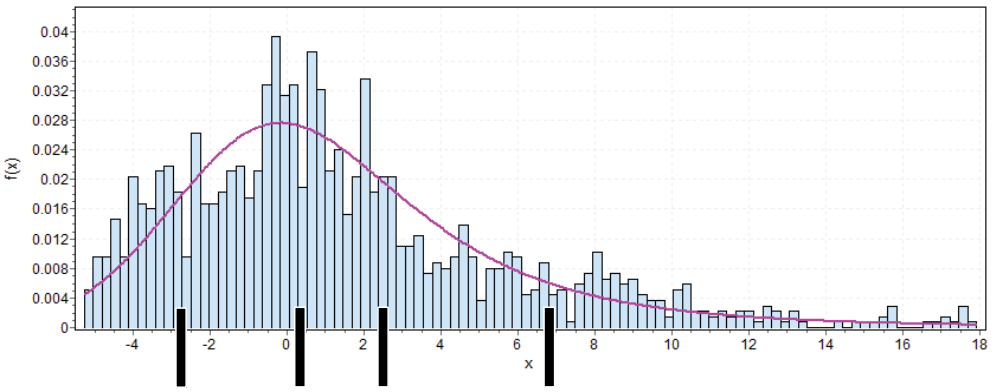
U odnosu na raspodelu podataka baze Banknote Authentication koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 82 - 85).



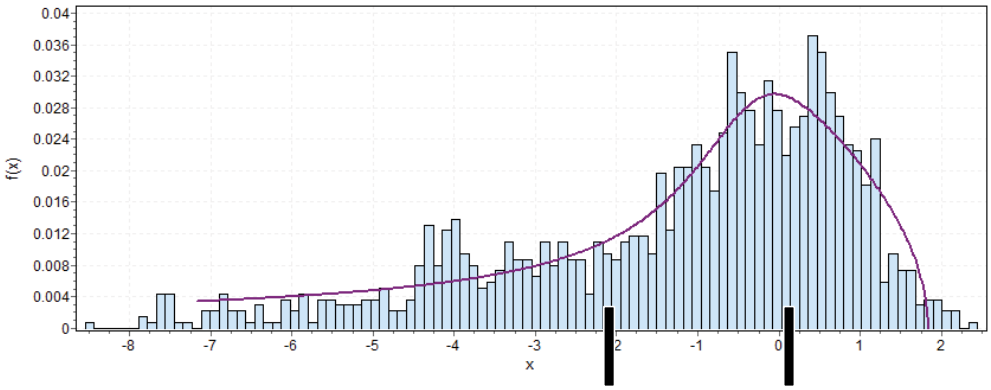
Slika 82. Raspodela podataka 1. atributa variance (baze Banknote Authentication) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 83. Raspodela podataka 2. atributa skewness (baze Banknote Authentication) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



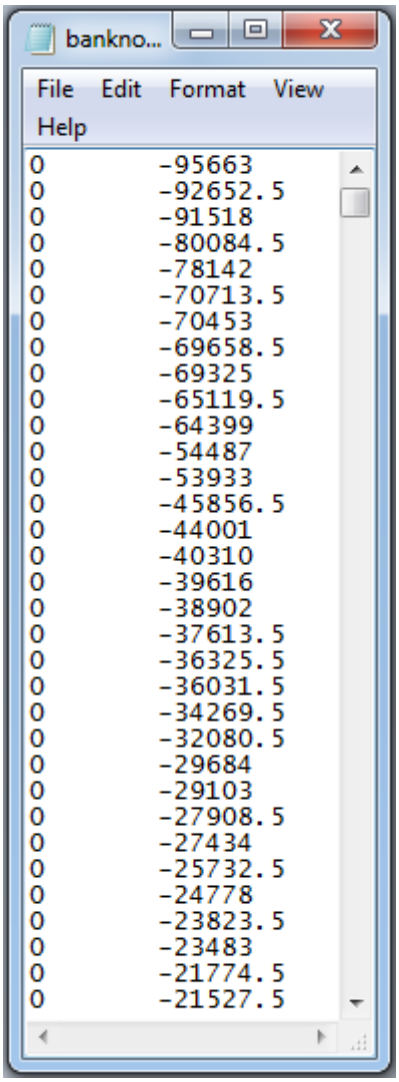
Slika 84. Raspodela podataka 3. atributa curtosis (baze Banknote Authentication) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 85. Raspodela podataka 4. atributa entropy (baze Banknote Authentication) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

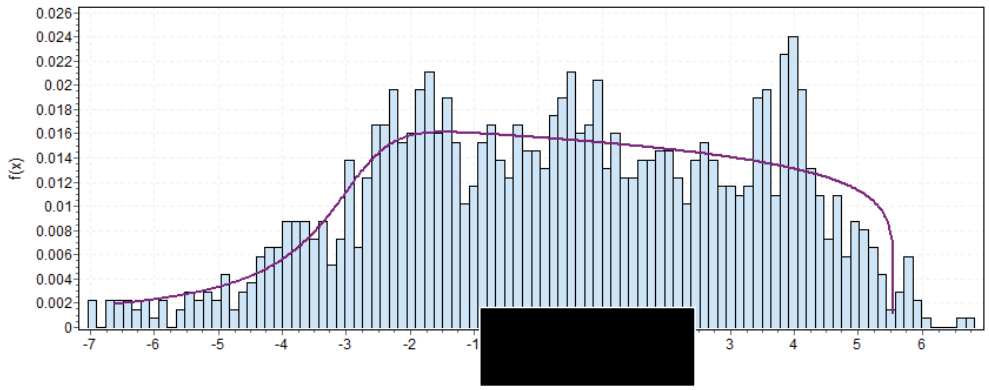
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Banknote Authentication algoritmom baziranim na entropiji, dobijene su tačke reza, čiji deo je prikazan na slici 86.

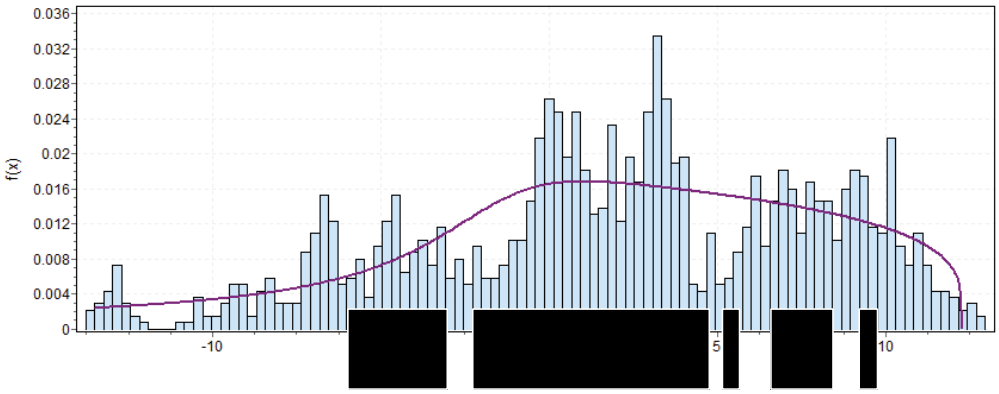


Slika 86. Deo tačaka reza baze Banknote Authentication dobijene algoritmom baziranim na entropiji

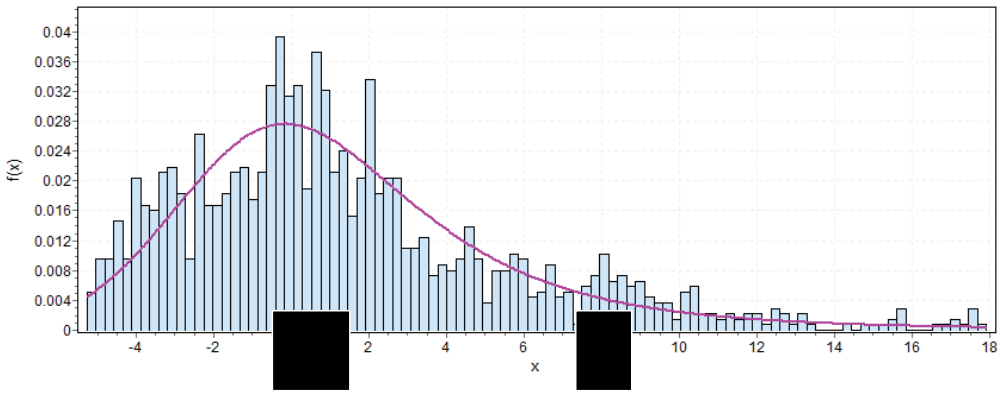
U odnosu na raspodelu podataka baze Banknote Authentication koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 87 - 90). Za prvi atribut variance postoji 174 tačke reza, za drugi atribut skewness postoji 236 tačaka reza, za treći atribut postoji 101 tačka reza, a za četvrti atribut postoji 309 tačaka reza.



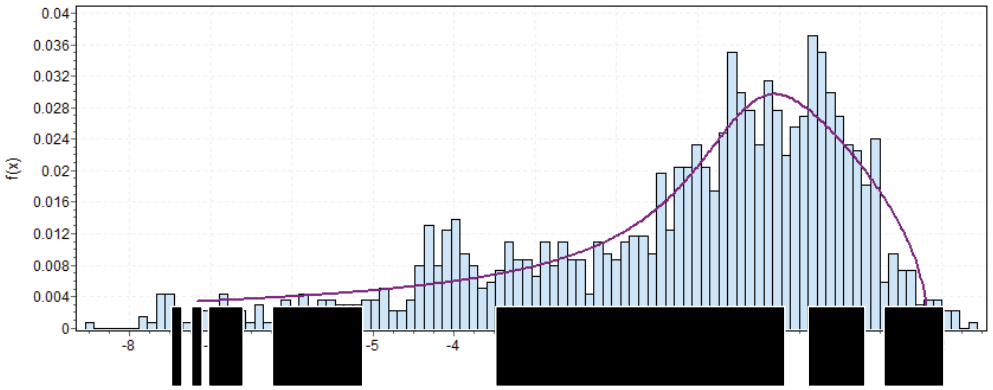
Slika 87. Raspodela podataka 1. atributa variance (baze Banknote Authentication) sa 174 tačke reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 88. Raspodela podataka 2. atributa skewness (baze Banknote Authentication) sa 236 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 89. Raspodela podataka 3. atributa curtosis (baze Banknote Authentication) sa 101 tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



Slika 90. Raspodela podataka 4. atributa entropy (baze Banknote Authentication) sa 309 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji

4. DODACI ZA ANALIZU BAZE GLASS

Baza Glass Identification potiče od USA Forensic Science Service a donor Vina Spiehler je uradio istraživanje za procenu 6 vrsta stakala u smislu njihovog oksid sadržaja (Na, Fe, K I ostalo).

Informacije o atributima (preuzeto iz [Glas, 2014]):

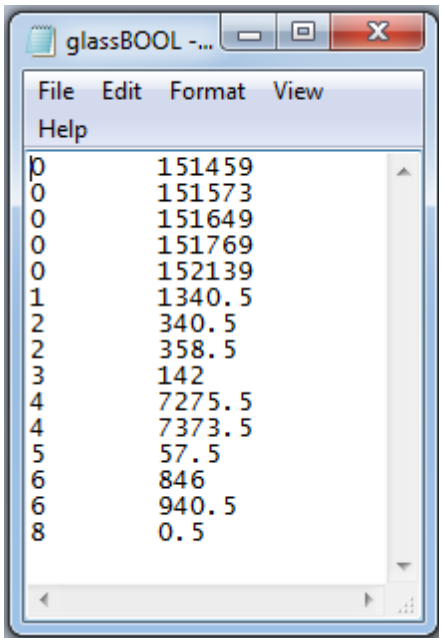
1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute)
-- 1 building_windows_float_processed
-- 2 building_windows_non_float_processed
-- 3 vehicle_windows_float_processed
-- 4 vehicle_windows_non_float_processed (none in this database)
-- 5 containers
-- 6 tableware
-- 7 headlamps

Broj instanci je 214. Prvi atribut je Id broj tako da je on u samom početku obrade podataka izbačen. Zbog toga je dalje analizirano 9 uslovnih atributa. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

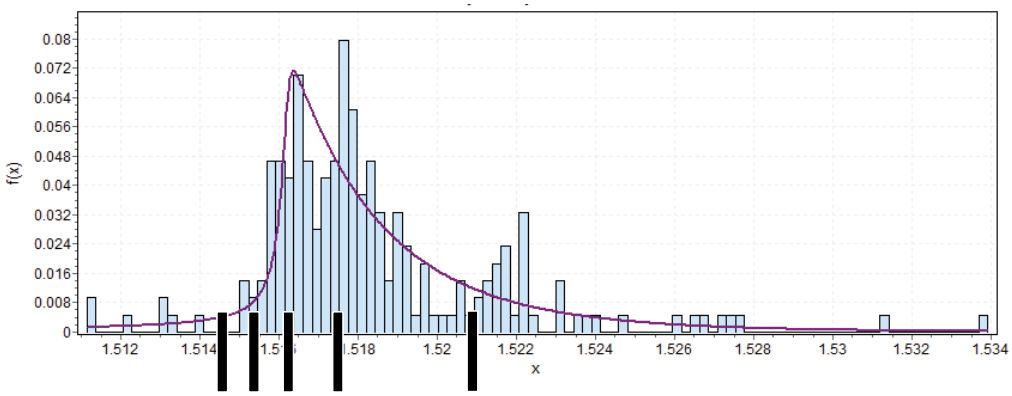
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Glass Identification dobijene su tačke reza koje su prikazane na slici 91.

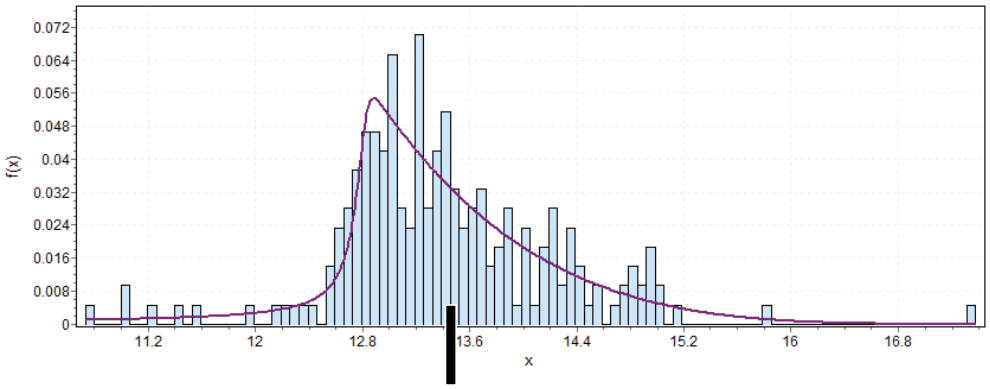
U odnosu na raspodelu podataka baze Glass Identification koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 92 - 100).



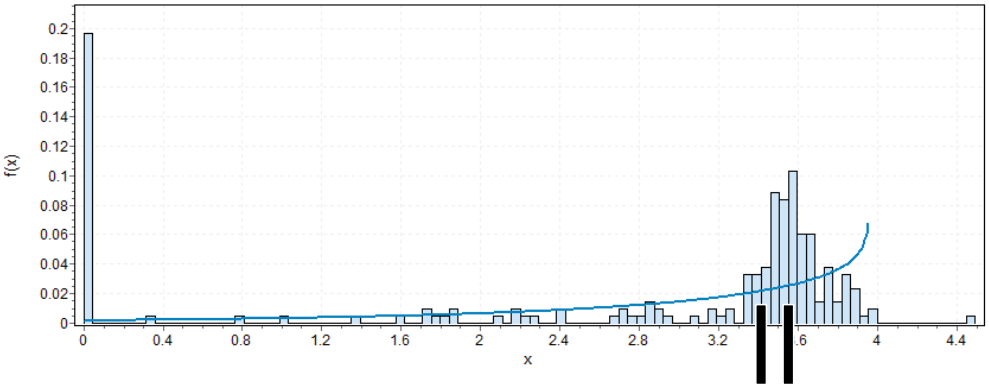
Slika 91. Tačke reza baze Glass Identification dobijene algoritmom maksimalne razberivosti



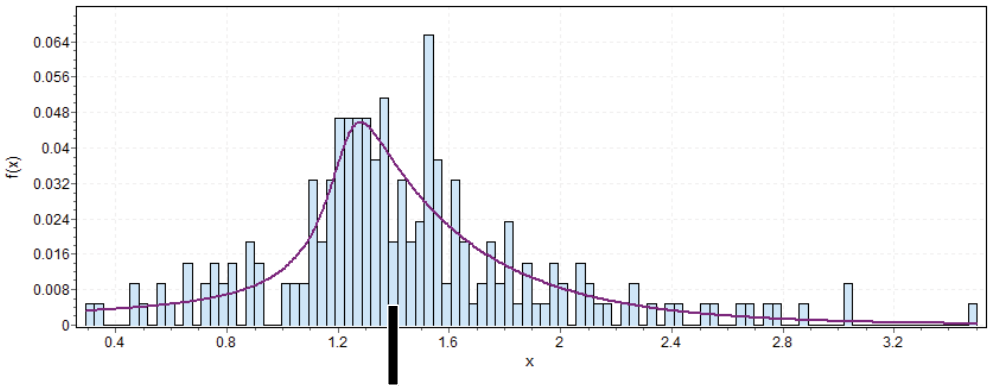
Slika 92. Raspodela podataka 2. atributa RI (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



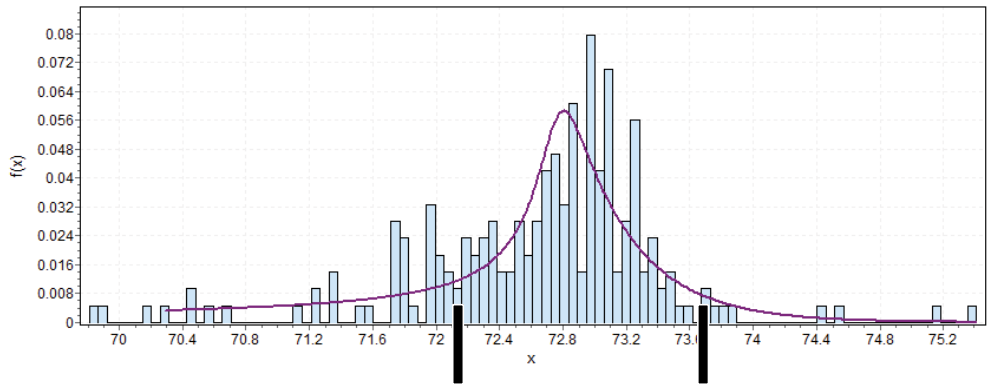
Slika 93. Raspodela podataka 3. atributa Na (baze Glass Identification) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



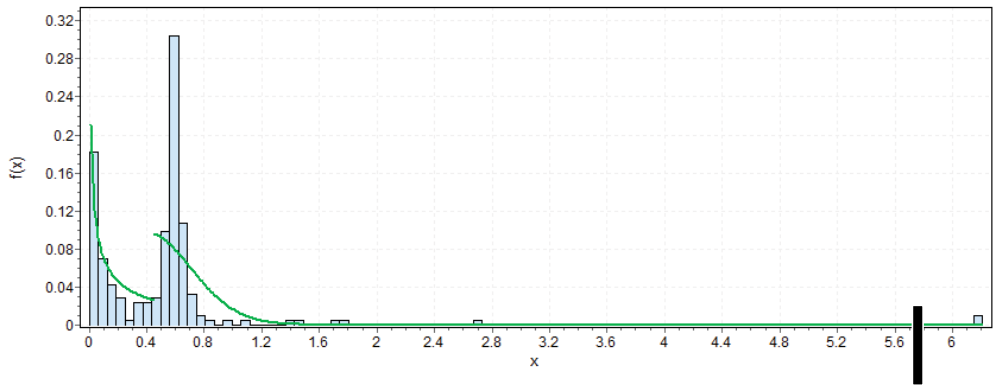
Slika 94. Raspodela podataka 4. atributa Mg (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



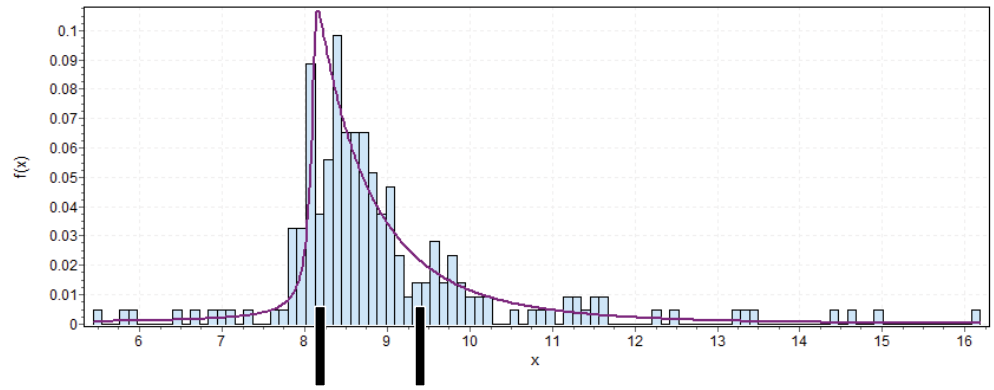
Slika 95. Raspodela podataka 5. atributa Al (baze Glass Identification) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



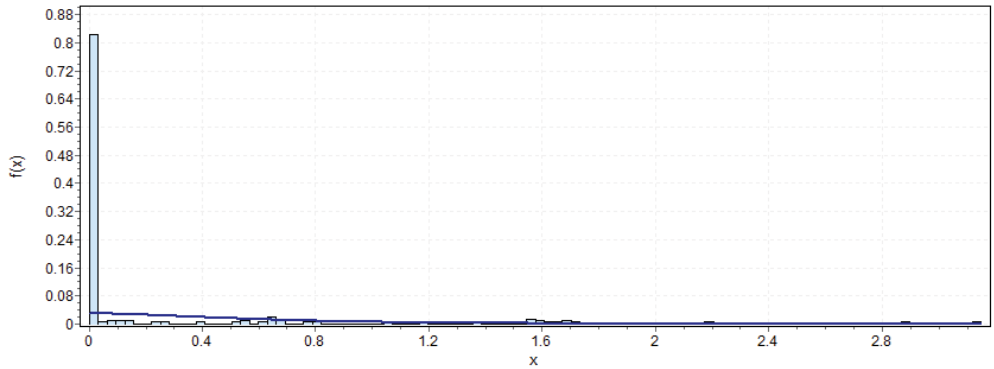
Slika 96. Raspodela podataka 6. atributa Si (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



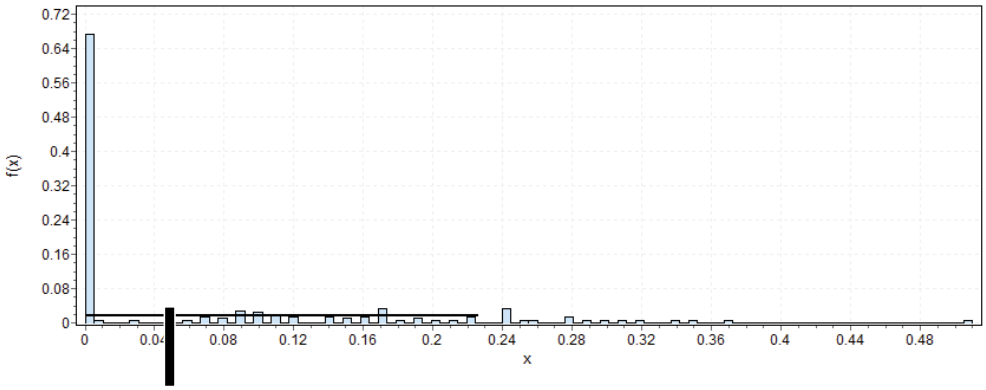
Slika 97. Raspodela podataka 7. atributa K (baze Glass Identification) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



Slika 98. Raspodela podataka 8. atributa Ca (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



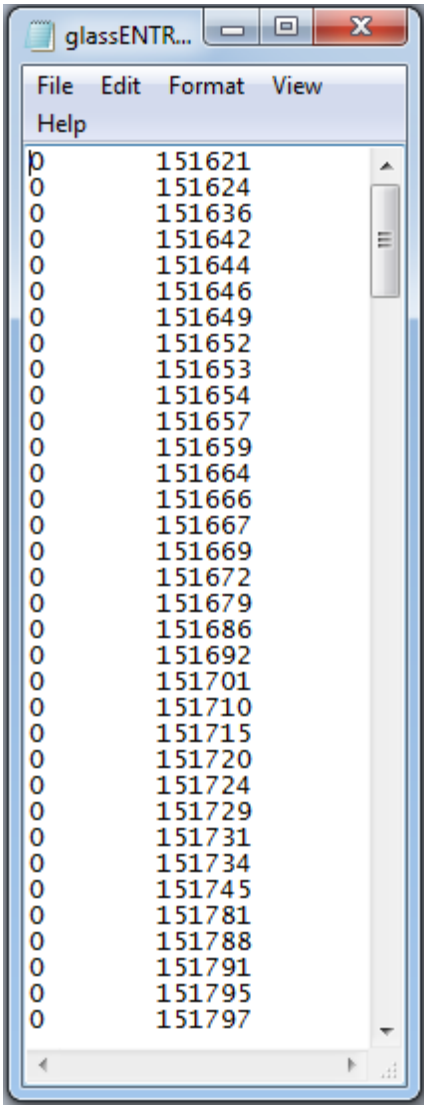
Slika 99. Raspodela podataka 9. atributa Ba (baze Glass Identification) nema tačke reza na osnovu algoritma maksimalne razberivosti



Slika 100. Raspodela podataka 10. atributa Fe (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

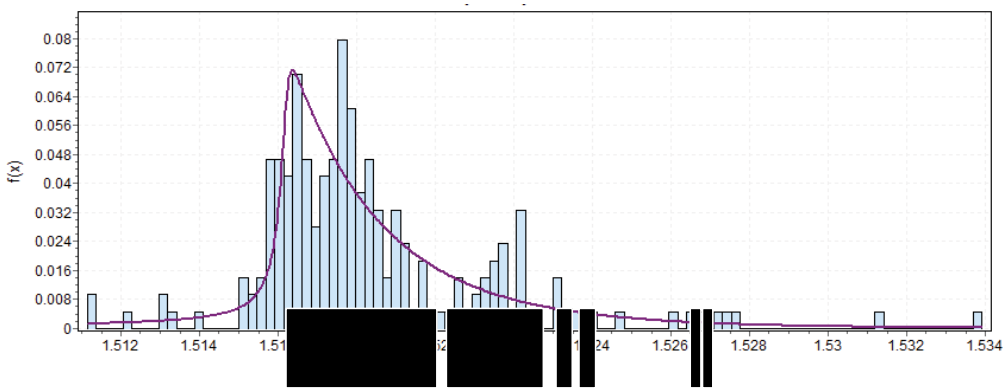
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Glass Identification algoritmom baziranim na entropiji, dobijene su tačke reza, čiji deo je prikazan na slici 101.

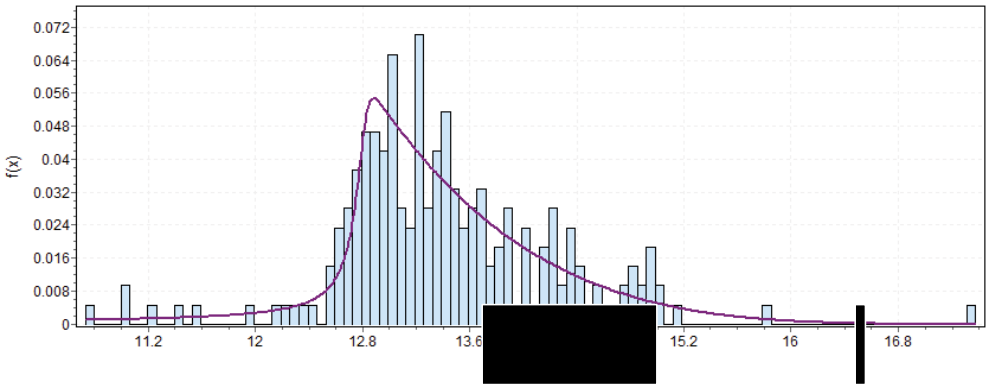


Slika 101. Deo tačaka reza baze Glass Identification dobijene algoritmom baziranim na entropiji

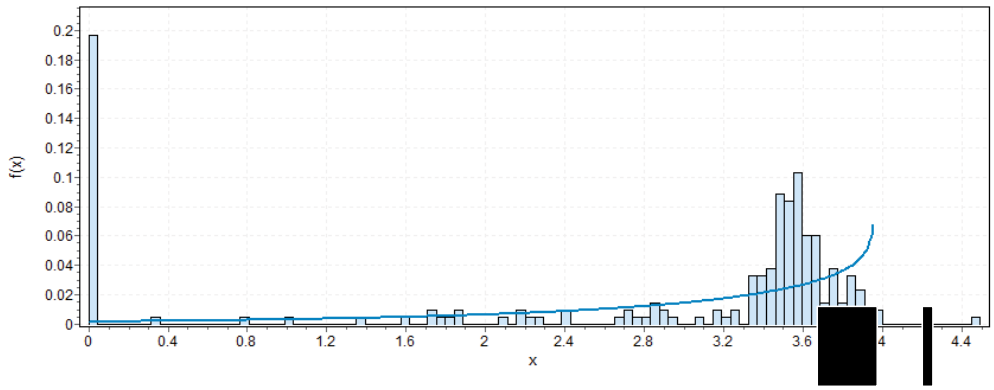
U odnosu na raspodelu podataka baze Glass Identification koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 102 - 110). Za analizirane atribute postoji značajno veći broj tačaka reza, tako da za drugi atribut RI postoji 87 tačaka reza, za treći atribut Na postoji 36 tačaka reza, za četvrti atribut Mg postoji 13 tačaka reza, za peti atribut Al postoji 25 tačaka reza, za šesti atribut Si postoji 17 tačaka reza, za sedmi atribut K postoji 11 tačaka reza, za osmi atribut Ca postoji 40. tačaka reza, za deveti atribut Ba postoje 3 tačke reza i za deseti atribut Fe postoji 8 tačaka reza.



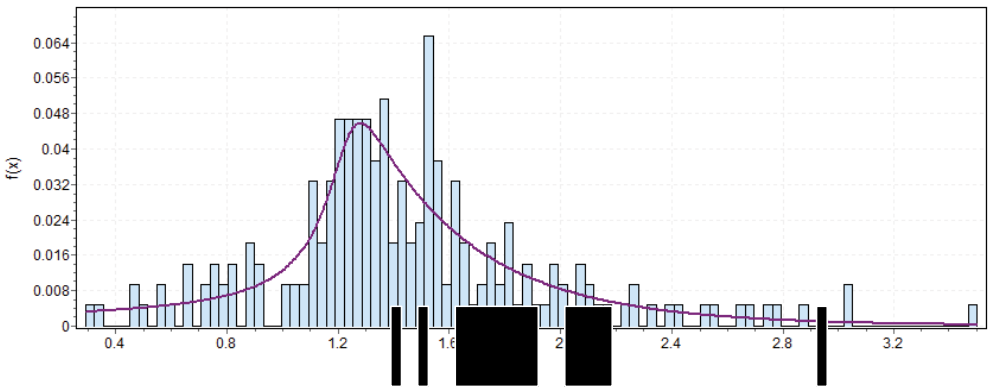
Slika 102. Raspodela podataka 2. atributa RI (baze Glass Identification) sa 87 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



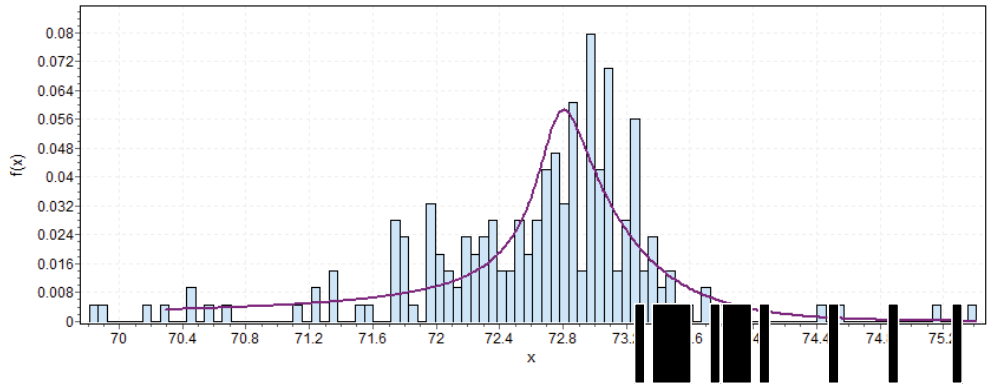
Slika 103. Raspodela podataka 3. atributa Na (baze Glass Identification) sa 36 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



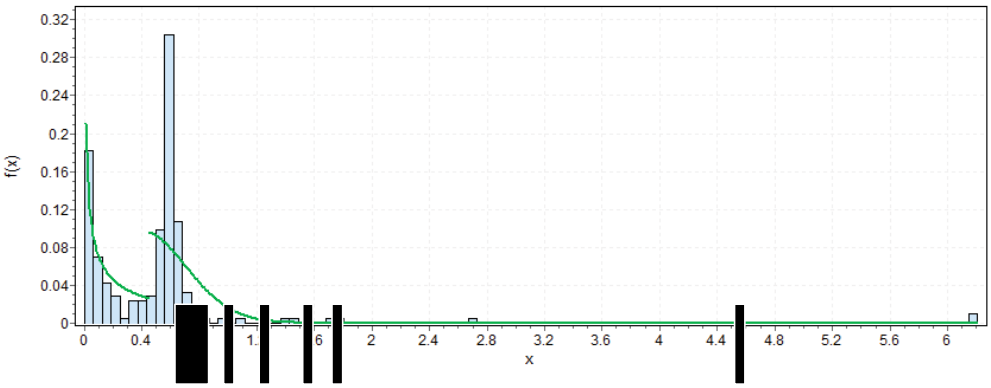
Slika 104. Raspodela podataka 4. atributa Mg (baze Glass Identification) sa 13 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



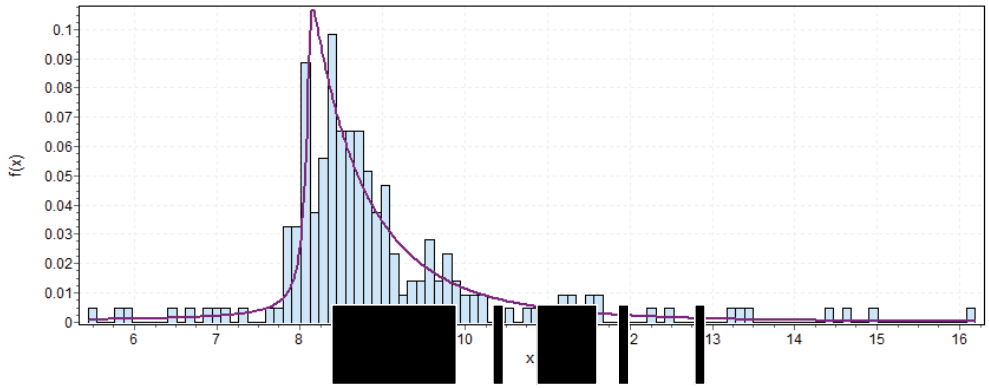
Slika 105. Raspodela podataka 5. atributa Al (baze Glass Identification) sa 25 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



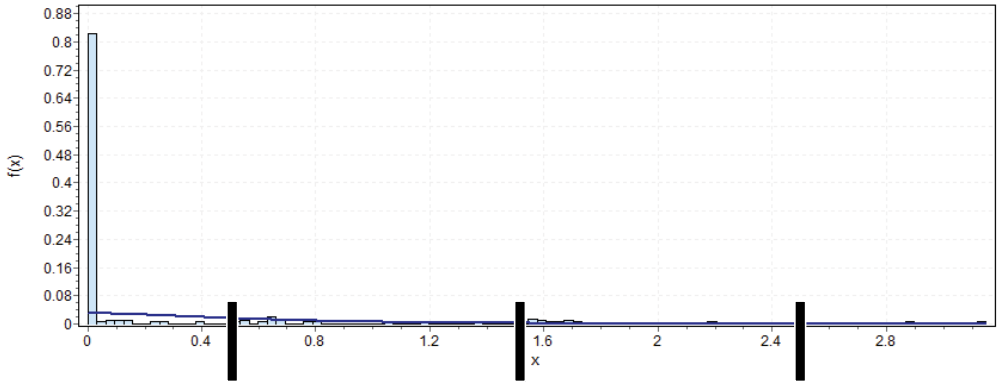
Slika 106. Raspodela podataka 6. atributa Si (baze Glass Identification) sa 17 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



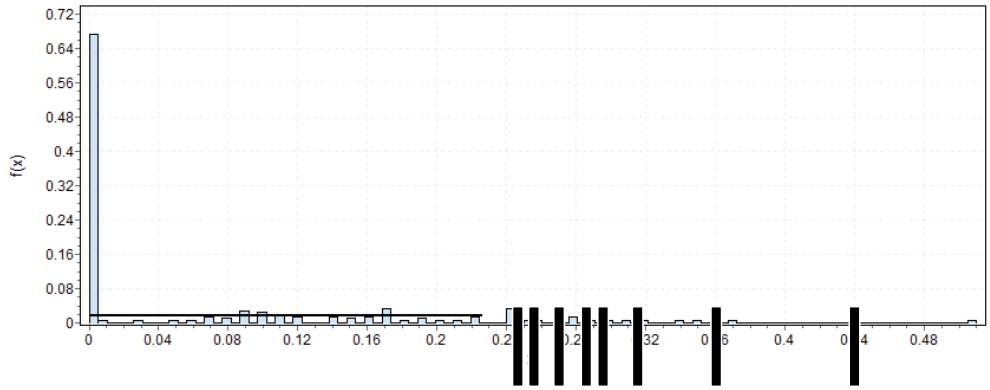
Slika 107. Raspodela podataka 7. atributa K (baze Glass Identification) sa 11 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 108. Raspodela podataka 8. atributa Ca (baze Glass Identification) sa 40 tačaka reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 109. Raspodela podataka 9. atributa Ba (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 110. Raspodela podataka 10. atributa Fe (baze Glass Identification) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji

5. DODACI ZA ANALIZU BAZE WILT DATA SET

Bazu je donirao Brian Johnson sa Institute for Global Environmental Strategies, Kanagawa, Japan [Wilt, 2015]. Ona sadrži podatke iz studije daljinskog očitavanja koja je obuhvatala detektovanje bolesnih drva pomoću Quickbird imagery. Baza se sastoji od segmenata slika koje su generisane segmentisanjem pan-izoštrene slike. Njeni segmenti sadrže informacije o spektru iz Quickbird multispektralnih slikovnih grupa i informacije o teksturi iz panhromatskih slikovnih grupa. Korišćen je deo baze za trening sa 4339 instanci.

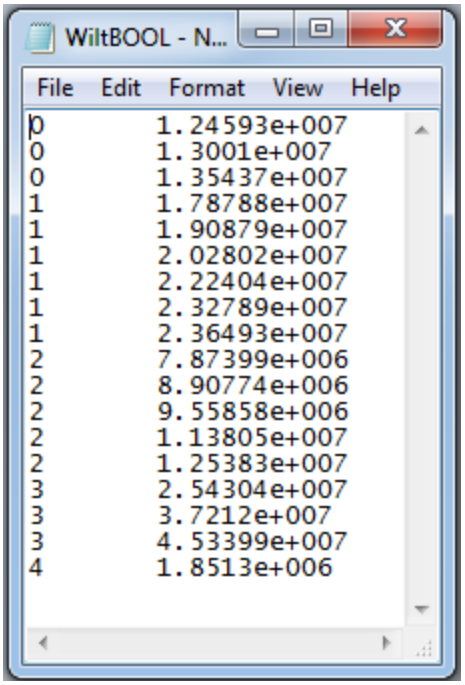
Informacije o atributima (preuzeto iz [Wilt, 2015]):

class: 'w' (diseased trees), 'n' (all other land cover)
GLCM_Pan: GLCM mean texture (Pan band)
Mean_G: Mean green value
Mean_R: Mean red value
Mean_NIR: Mean NIR value
SD_Pan: Standard deviation (Pan band)

Broj instanci je 4339. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

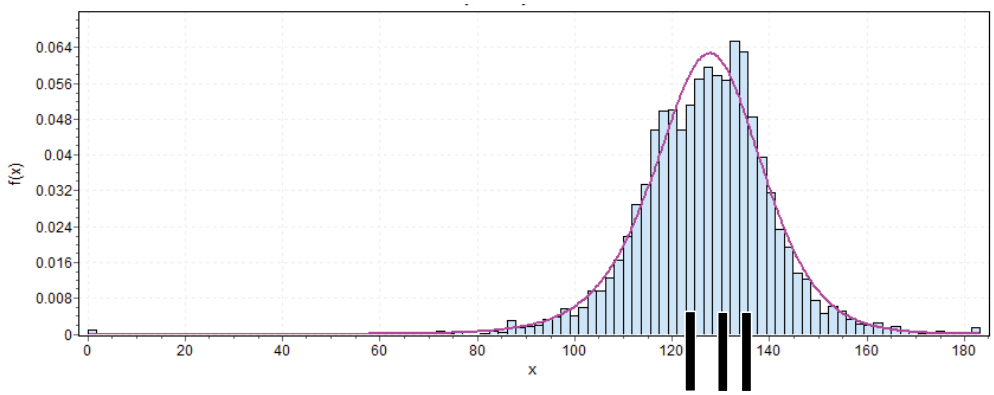
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Wilt Data Set algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 111.

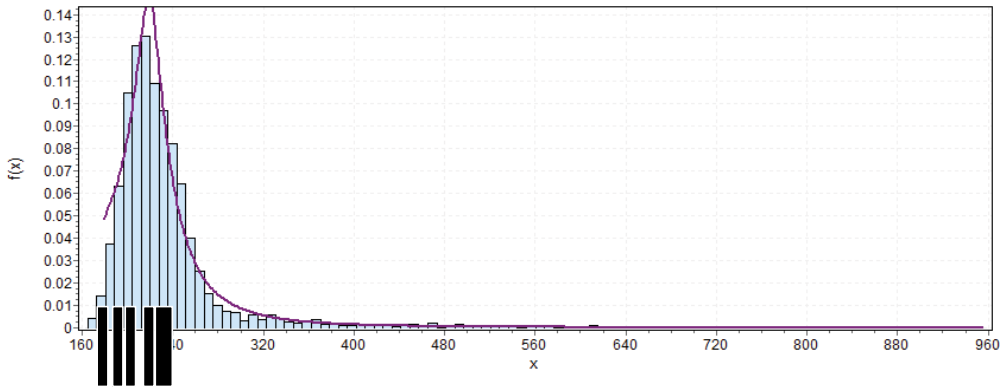


Slika 111. Tačke reza baze Wilt Data Set dobijene algoritmom maksimalne razberivosti

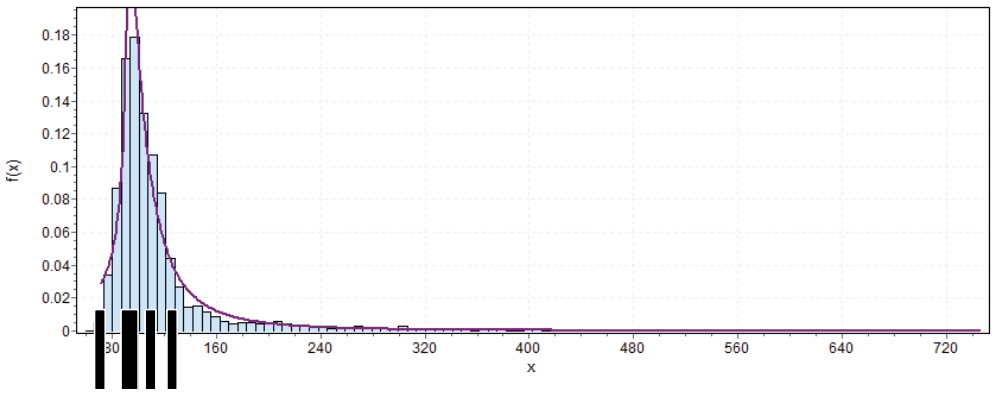
U odnosu na raspodelu podataka baze Wilt Data Set koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama na slikama 112. do 116.



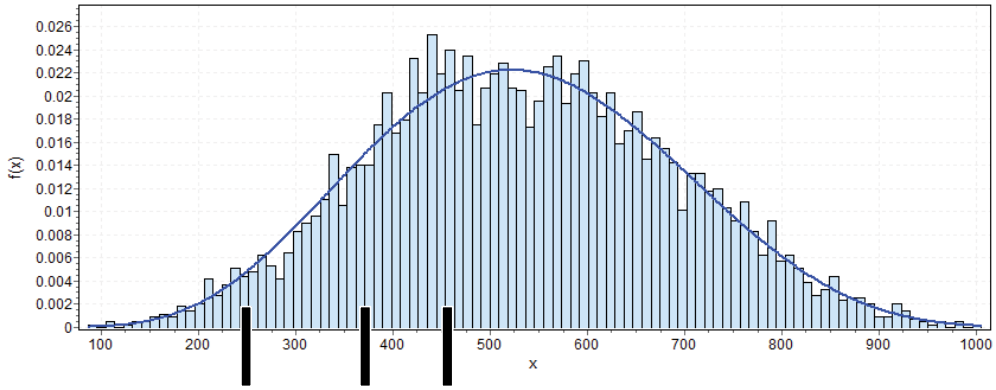
Slika 112. Raspodela podataka 1. atributa GLCM_Pan (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



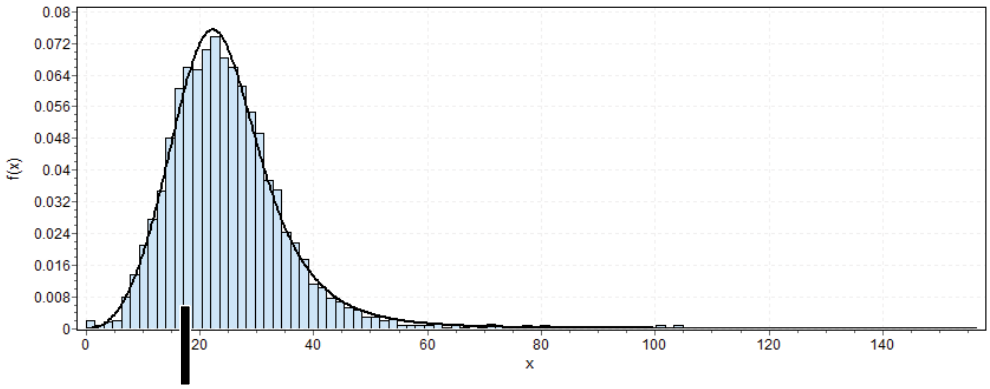
Slika 113. Raspodela podataka 2. atributa Mean_G (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 114. Raspodela podataka 3. atributa Mean_R (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



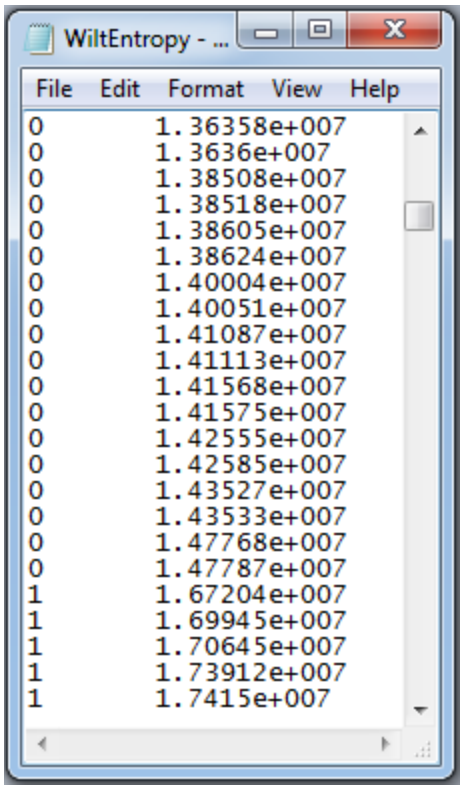
Slika 115. Raspodela podataka 4. atributa Mean_NIR (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 116. Raspodela podataka 5. atributa SD_Pan (baze Wilt Data Set) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti

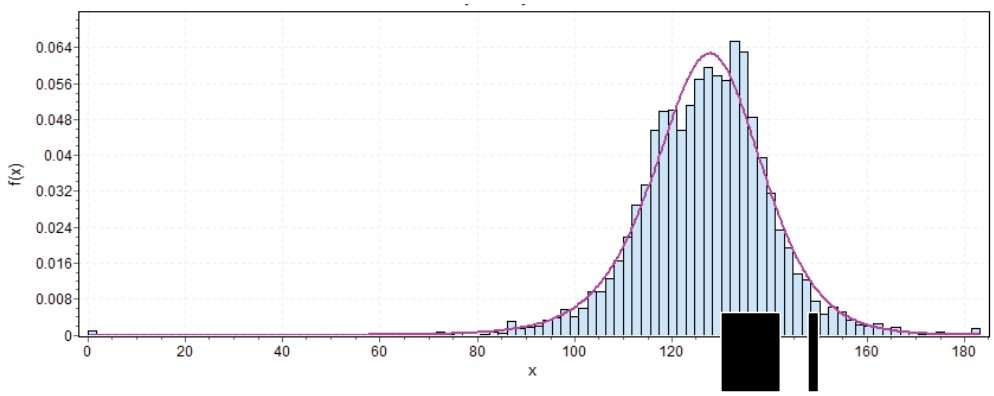
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Wilt Data Set algoritmom baziranim na entropiji, dobijene su tačke reza, a deo njih je prikazan na slici 117. Prvi atribut ima 55 tačaka reza, drugi atribut ima 69 tačaka reza, treći atribut ima 19 tačaka reza, četvrti atribut ima 97 tačaka reza i peti atribut ima 145 tačaka reza.

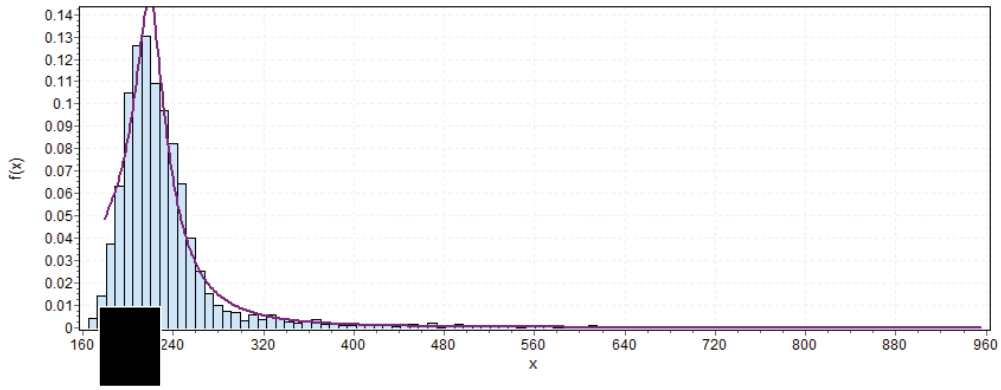


Slika 117. Deo tačaka reza baze Wilt Data Set dobijene algoritmom baziranim na entropiji

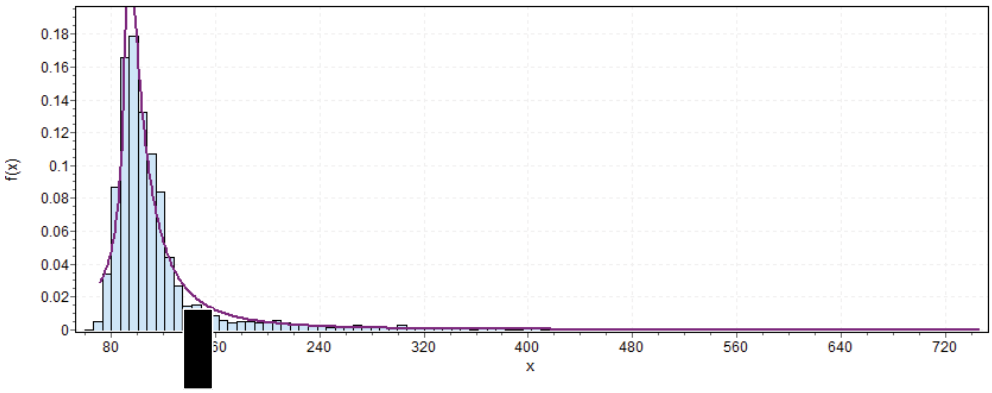
U odnosu na raspodelu podataka baze Wilt Data Set koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 118 – 122).



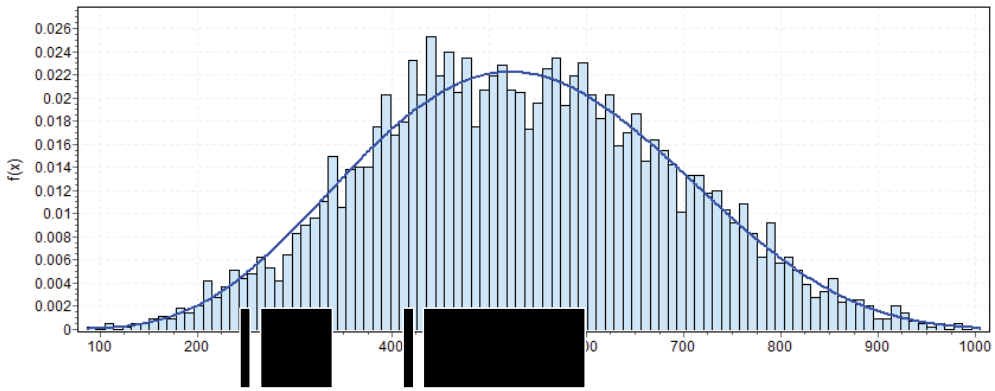
Slika 118. Raspodela podataka 1. atributa GLCM_Pan (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



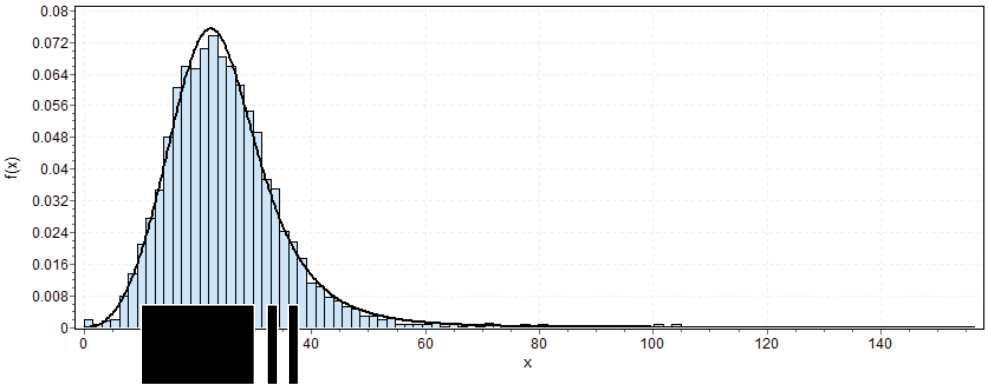
Slika 119. Raspodela podataka 2. atributa 2. atributa Mean_G (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 120. Raspodela podataka 3. atributa Mean_R (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 121. Raspodela podataka 4. atributa Mean_NIR (baze Wilt Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 122. Raspodela podataka 5. atributa SD_Pan (baze Wilt Data Set) bez tačke reza na osnovu algoritma baziranog na entropiji

6. DODACI ZA ANALIZU BAZE BREAST CANCER WISCONSIN DATA SET

Bazu Breast Cancer Wisconsin je kreirao Dr. WIlliam H. Wolberg, University of Wisconsin Hospitals, Madison, Wisconsin, USA.

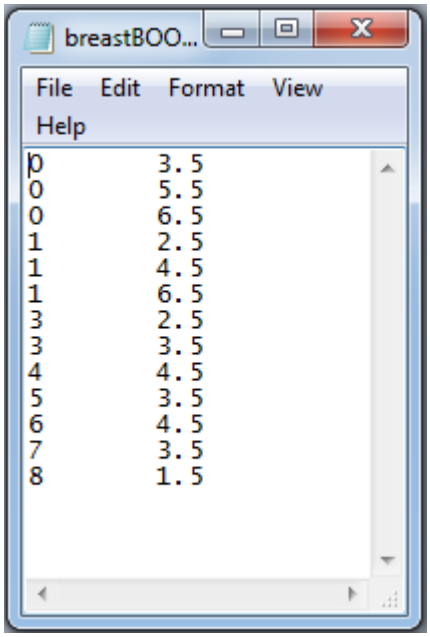
Informacije o atributima (preuzeto iz [Breast, 2011]):

1. Sample code number: id number
2. Clump Thickness: 1 – 10
3. Uniformity of Cell Size: 1 – 10
4. Uniformity of Cell Shape: 1 – 10
5. Marginal Adhesion: 1 – 10
6. Single Epithelial Cell Size: 1 – 10
7. Bare Nuclei: 1 – 10
8. Bland Chromatin: 1 – 10
9. Normal Nucleoli: 1 – 10
10. Mitoses: 1 – 10
11. Class: (2 for benign, 4 for malignant)

Broj instanci je 699. Prvi atribut je Id broj tako da je on u samom početku obrade podataka izbačen. Zbog toga je dalje analizirano 9 uslovnih atributa. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

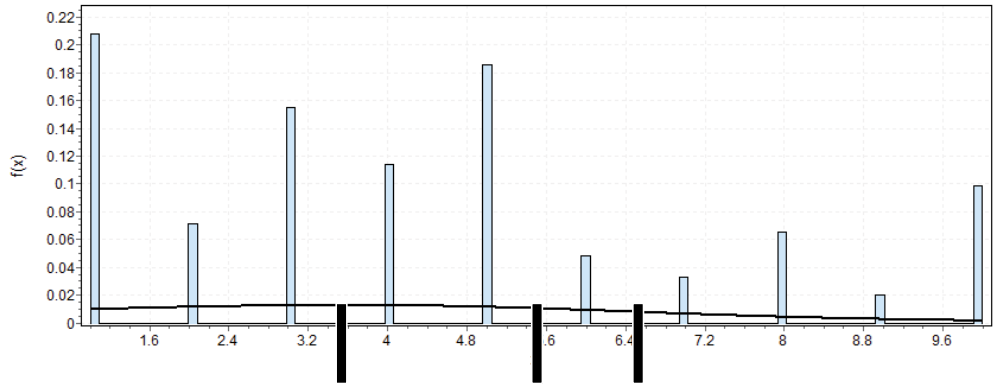
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Breast Cancer Wisconsin algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 123.

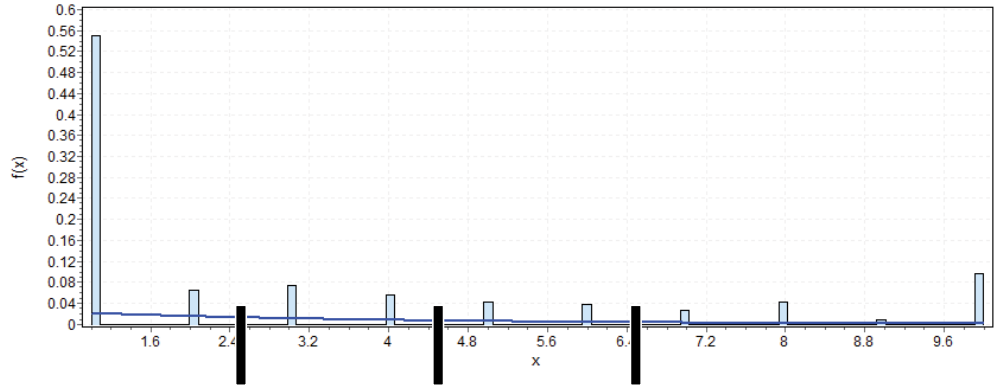


Slika 123. Tačke reza baze Breast Cancer Wisconsin dobijene algoritmom maksimalne razberivosti

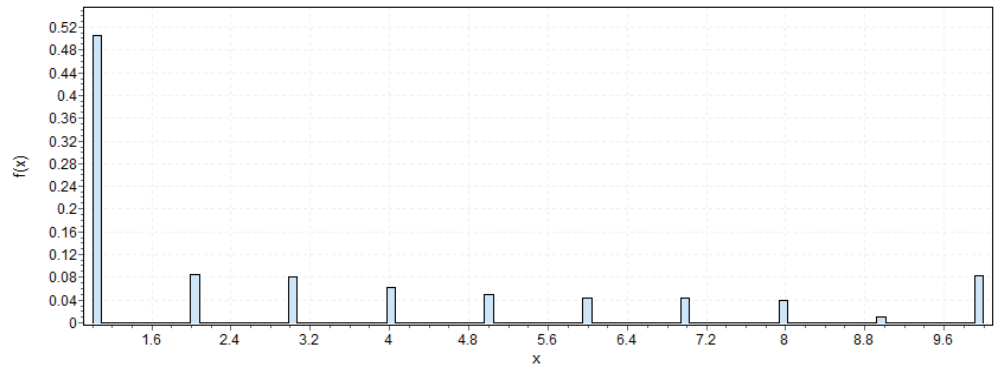
U odnosu na raspodelu podataka baze Breast Cancer Wisconsin koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 124 - 132).



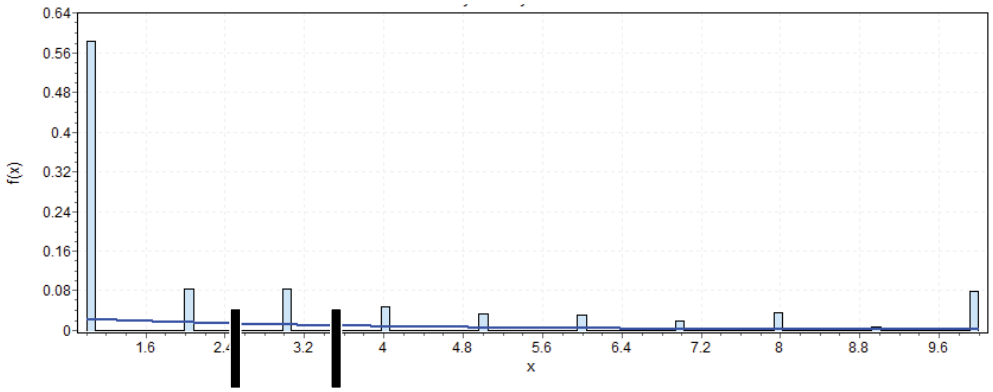
Slika 124. Raspodela podataka 2. atributa Clump Thickness (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



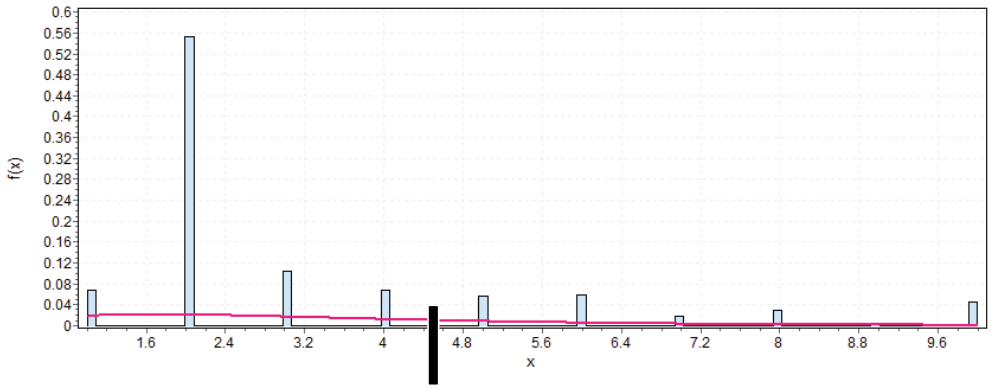
Slika 125. Raspodela podataka 3. atributa Uniformity of Cell Size (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



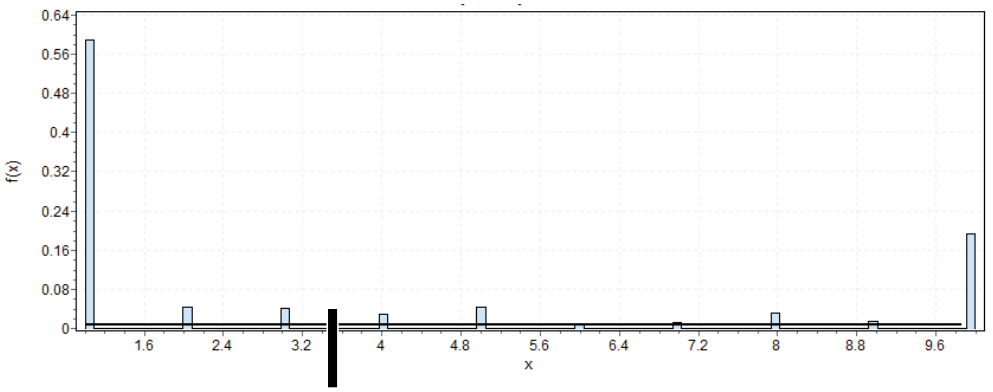
Slika 126. Raspodela podataka 4. atributa Uniformity of Cell Shape (baze Breast Cancer Wisconsin) bez tačke reza na osnovu algoritma maksimalne razberivosti



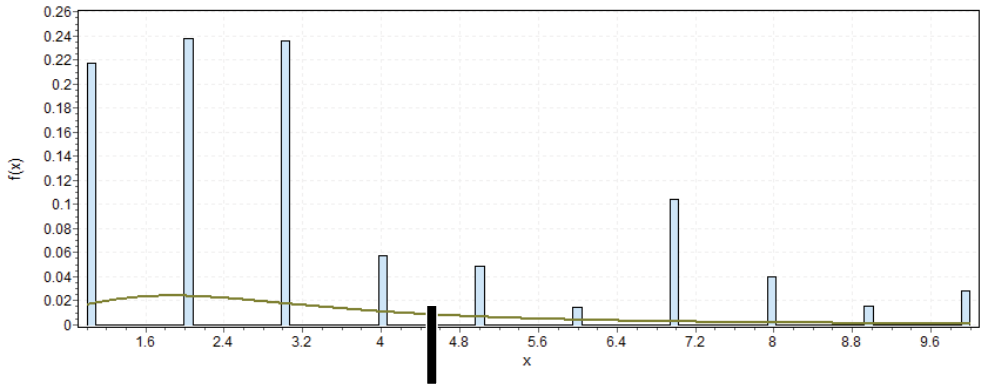
Slika 127. Raspodela podataka 5. atributa Marginal Adhesion (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



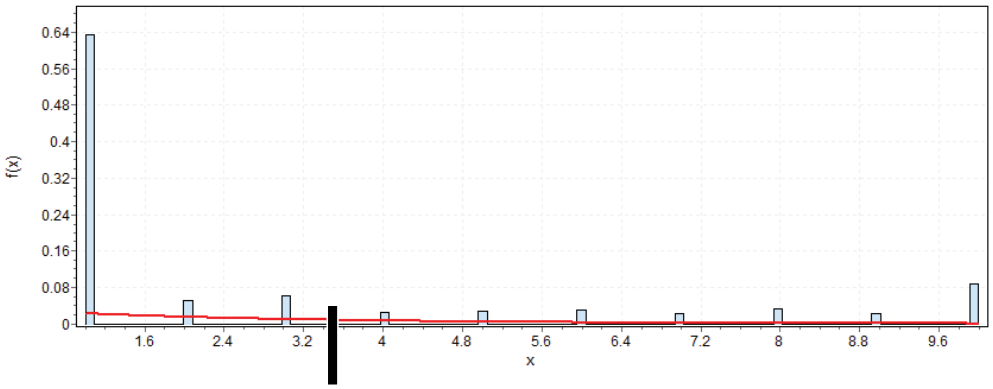
Slika 128. Raspodela podataka 6. atributa Single Epithelial Cell Size (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



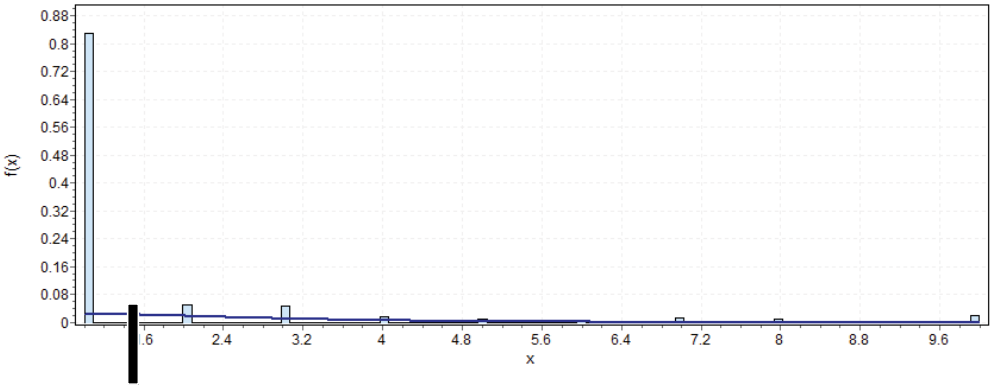
Slika 129. Raspodela podataka 7. atributa Bare Nuclei (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



Slika 130. Raspodela podataka 8. atributa Bland Chromatin (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



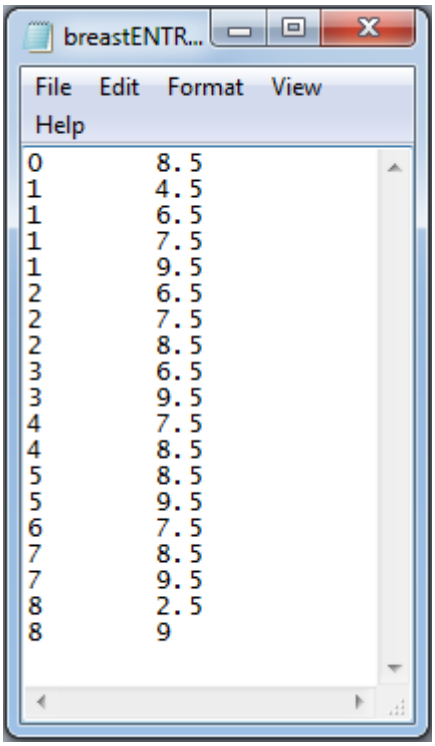
Slika 131. Raspodela podataka 9. atributa Normal Nucleoli (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



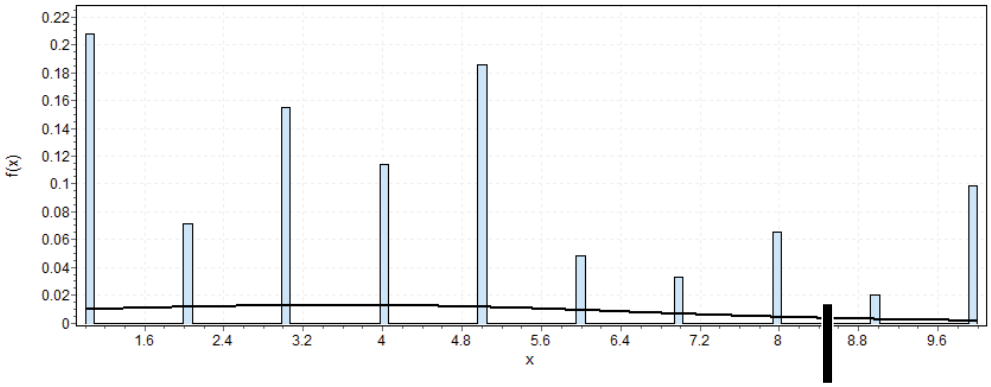
Slika 132. Raspodela podataka 10. atributa Mitoses (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti

Algoritam baziran na entropiji

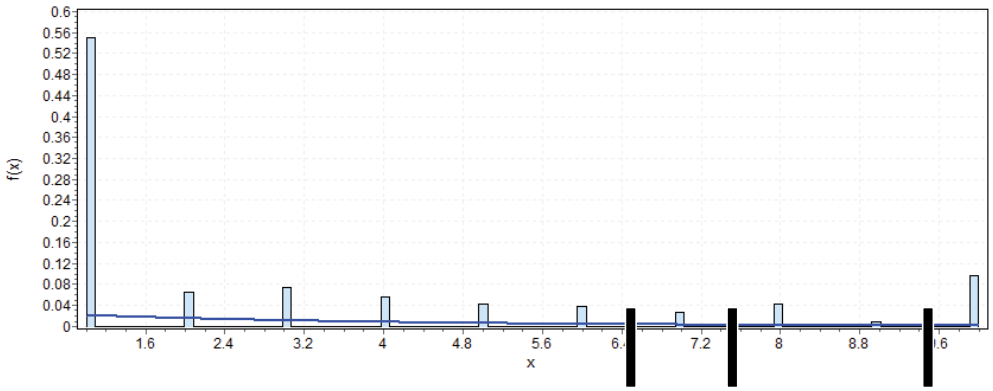
Na osnovu diskretizacije baze Breast Cancer Wisconsin algoritmom baziranim na entropiji, dobijene su tačke reza, čiji deo je prikazan na slici 133. U odnosu na raspodelu podataka baze Breast Cancer Wisconsin koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 134 - 142).



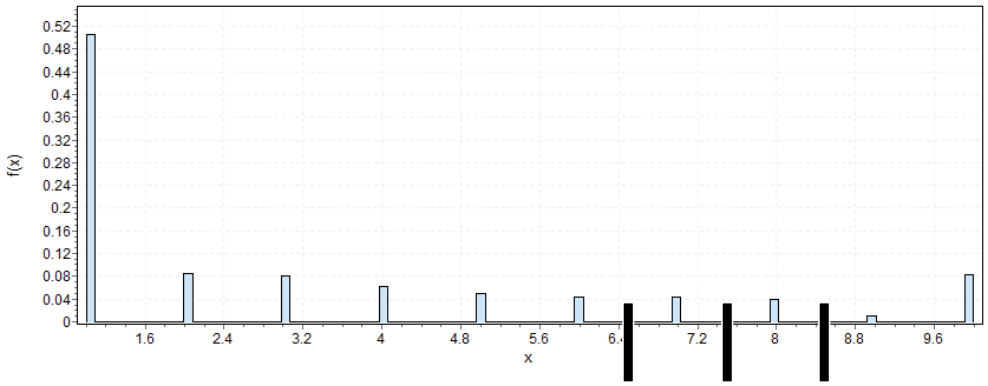
Slika 133. Tačke reza baze Breast Cancer Wisconsin dobijene algoritmom baziranim na entropiji



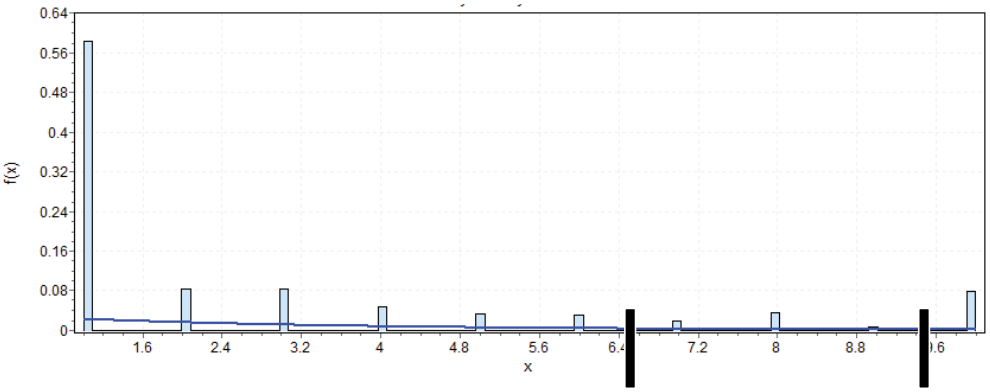
Slika 134. Raspodela podataka 2. atributa Clump Thickness (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



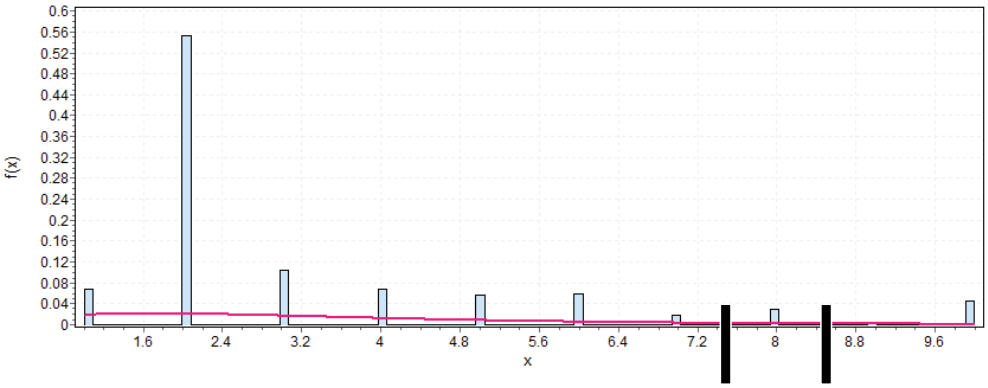
Slika 135. Raspodela podataka 3. atributa Uniformity of Cell Size (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



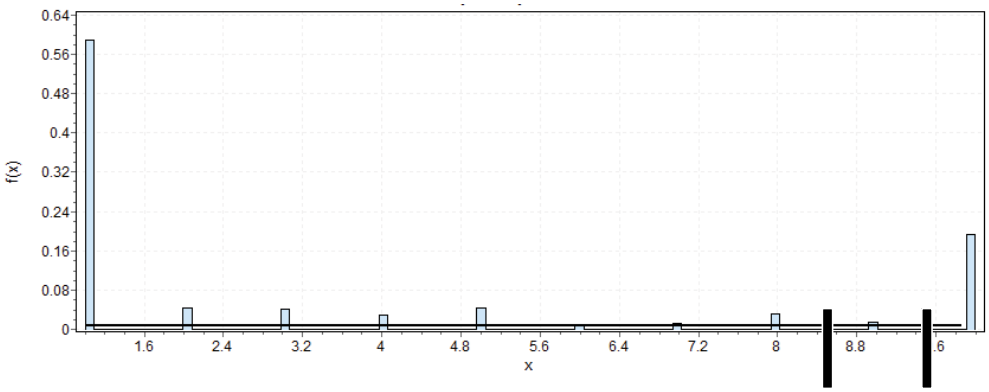
Slika 136. Raspodela podataka 4. atributa Uniformity of Cell Shape (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



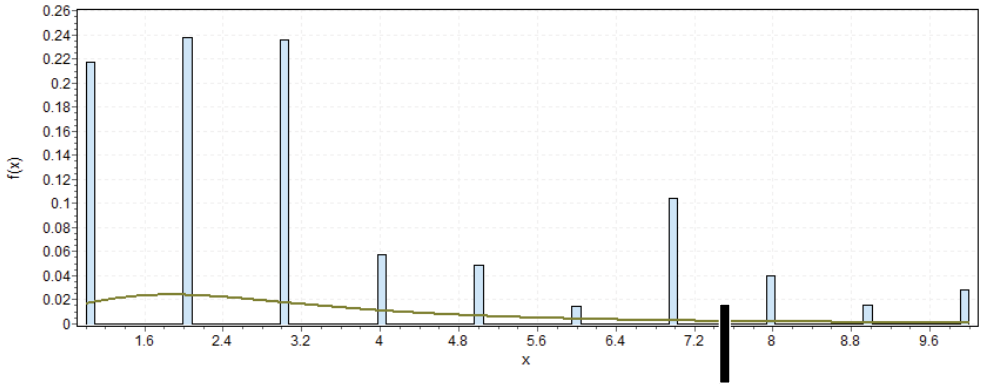
Slika 137. Raspodela podataka 5. atributa Marginal Adhesion (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



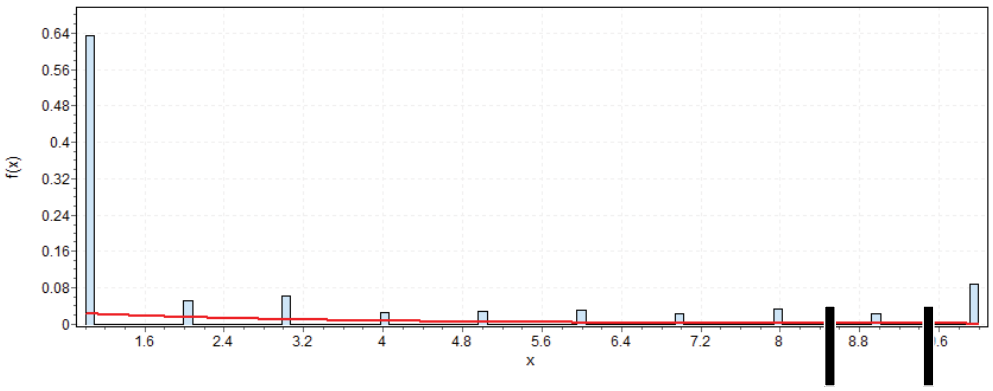
Slika 138. Raspodela podataka 6. atributa Single Epithelial Cell Size (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



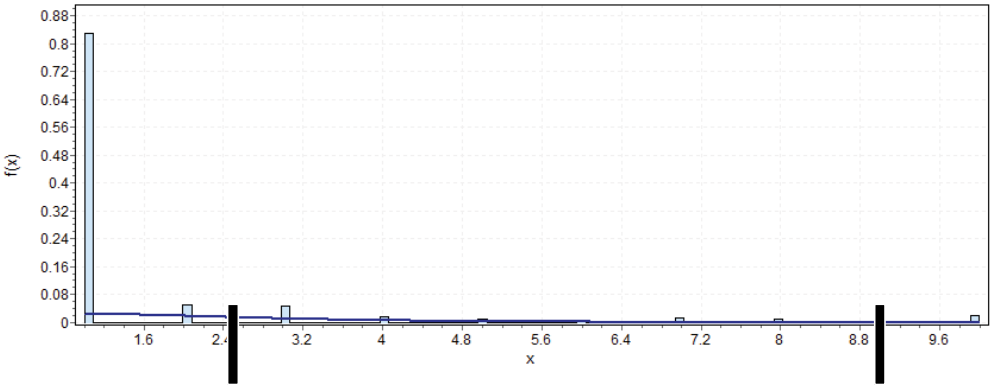
Slika 139. Raspodela podataka 7. atributa Bare Nuclei (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 140. Raspodela podataka 8. atributa Bland Chromatin (baze Breast Cancer Wisconsin) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



Slika 141. Raspodela podataka 9. atributa Normal Nucleoli (baze Breast Cancer Wisconsin) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 142. Raspodela podataka 10. atributa Mitoses (baze Breast Cancer Wisconsin) nema tačke reza na osnovu algoritma baziranog na entropiji

7. DODACI ZA ANALIZU BAZE CARDIOTOCOGRAPHY

Baza Cardiotocography potiče iz Biomedical Engineering Institute, Porto, Portugal. Ukupno 2126 fetalnih kardiotokograma (CTGa) su bili automatski procesuirani i izmerene su respektivne dijagnostičke odlike. CTG snimke su klasifikovala tri različita akušera i svakom od njih je dodeljena oznaka konsenzusa klasifikacije. Klasifikacija je takođe vršena vodeći računa o morfološkoj šemi (A, B, C,...) i fetalnom stanju (N, S, P). Prema tome, skup podataka može da se koristi i za desetoklasne i za troklasne eksperimente. Zbog toga će se 22. atribut koristiti kao uslovni.

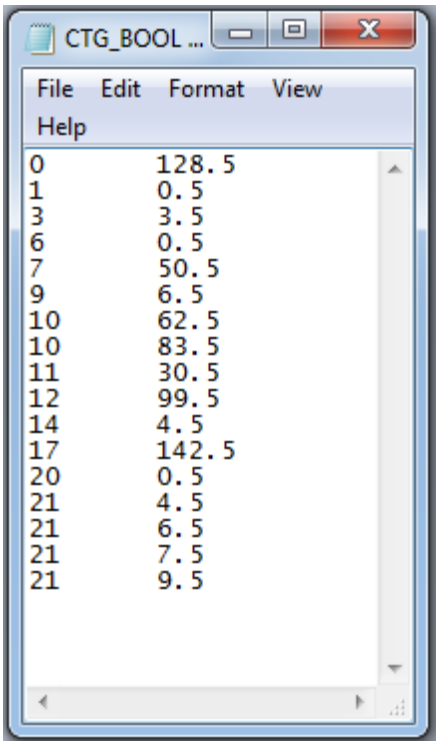
Informacije o atributima (preuzeto iz [Cardiotocography, 2015]):

1. LB - FHR baseline (beats per minute)
2. AC - # of accelerations per second
3. FM - # of fetal movements per second
4. UC - # of uterine contractions per second
5. DL - # of light decelerations per second
6. DS - # of severe decelerations per second
7. DP - # of prolongued decelerations per second
8. ASTV - percentage of time with abnormal short term variability
9. MSTV - mean value of short term variability
10. ALTV - percentage of time with abnormal long term variability
11. MLTV - mean value of long term variability
12. Width - width of FHR histogram
13. Min - minimum of FHR histogram
14. Max - Maximum of FHR histogram
15. Nmax - # of histogram peaks
16. Nzeros - # of histogram zeros
17. Mode - histogram mode
18. Mean - histogram mean
19. Median - histogram median
20. Variance - histogram variance
21. Tendency - histogram tendency
22. CLASS - FHR pattern class code (1 to 10)
23. NSP - fetal state class code (N=normal; S=suspect; P=pathologic)

Broj instanci je 2126. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

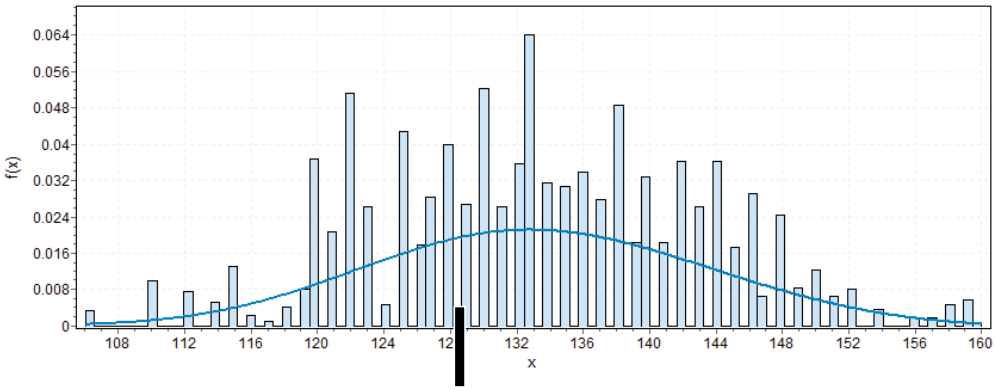
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Cardiotocography algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 143.

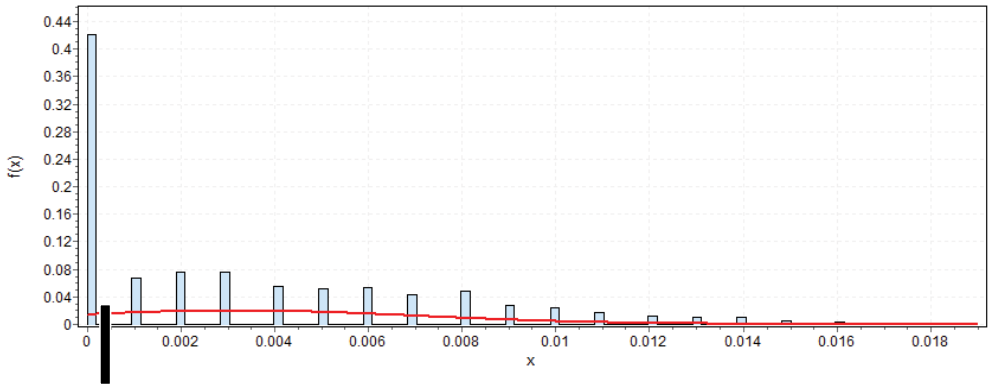


Slika 143. Tačke reza baze Cardiotocography dobijene algoritmom maksimalne razberivosti

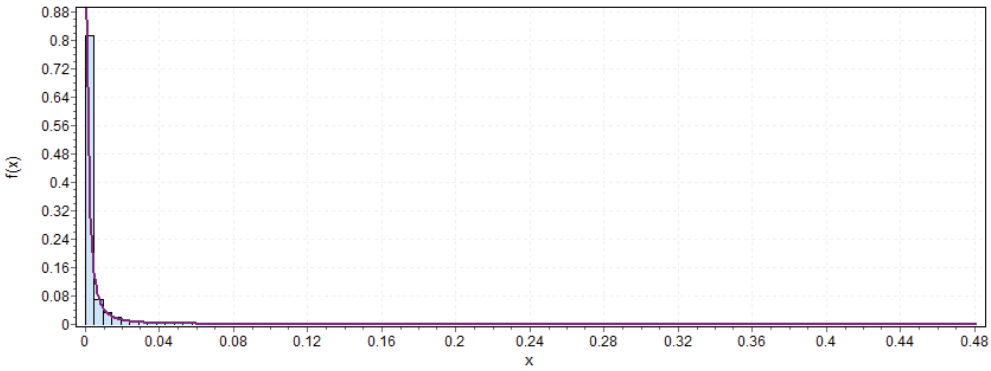
U odnosu na raspodelu podataka baze Cardiotocography koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 144 - 165).



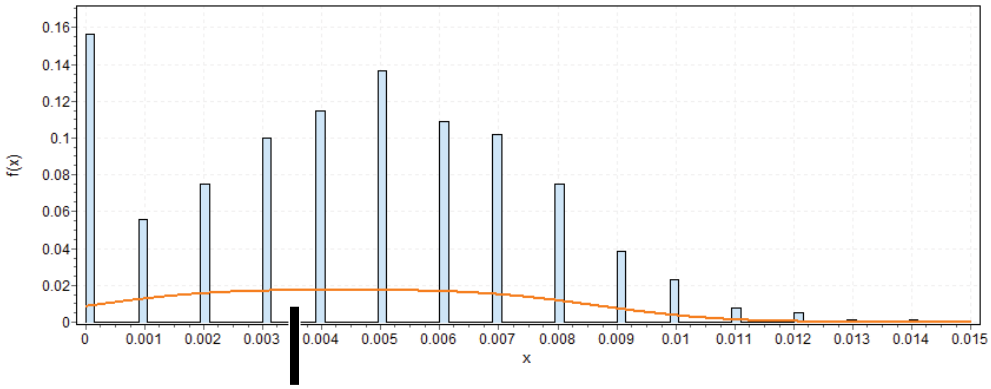
Slika 144. Raspodela podataka 1. atributa LB (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



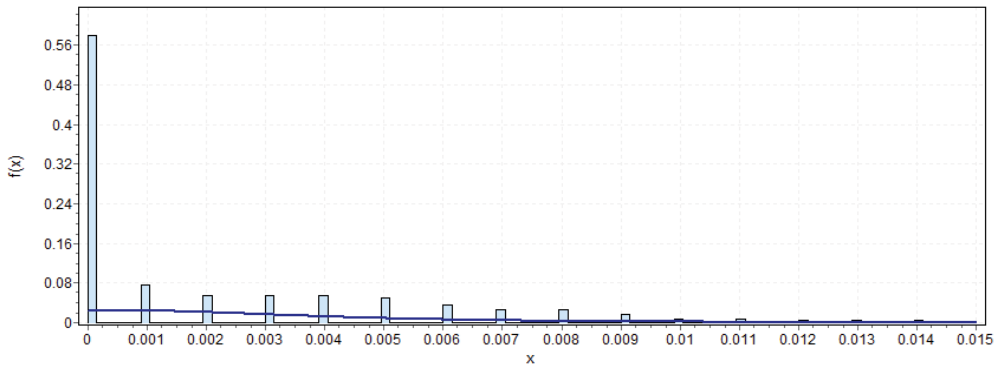
Slika 145. Raspodela podataka 2. atributa AC (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



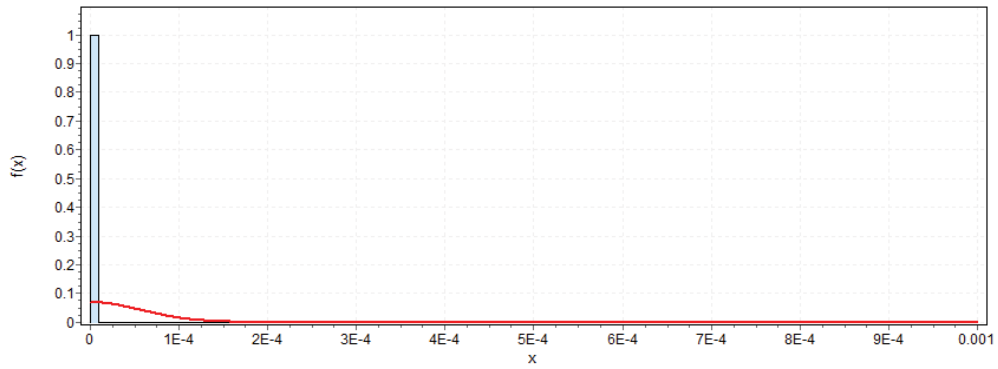
Slika 146. Raspodela podataka 3. atributa FM (baze Cardiotocography) nema tačaka reza na osnovu algoritma maksimalne razberivosti



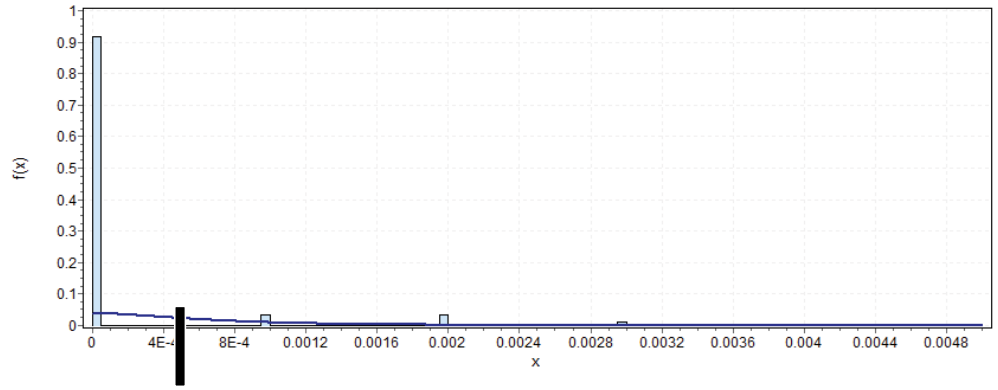
Slika 147. Raspodela podataka 4. atributa UC (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



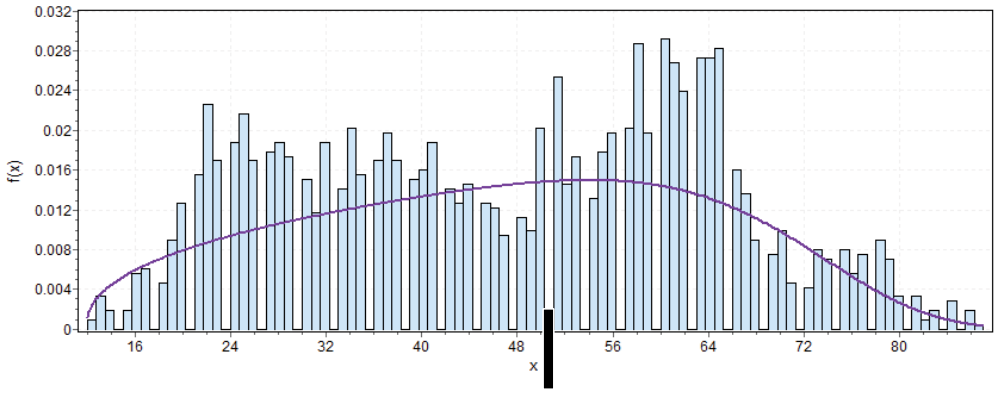
Slika 148. Raspodela podataka 5. atributa DL (baze Cardiotocography) nema tačaka reza na osnovu algoritma maksimalne razberivosti



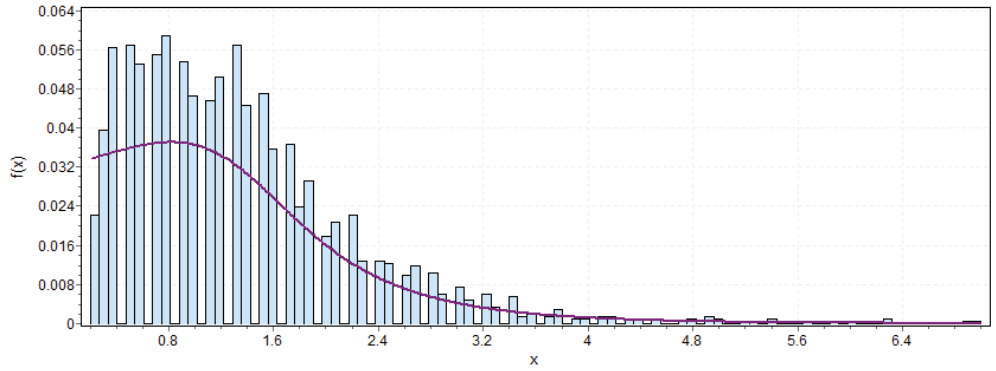
Slika 149. Raspodela podataka 6. atributa DS (baze Cardiotocography) nema tačaka reza na osnovu algoritma maksimalne razberivosti



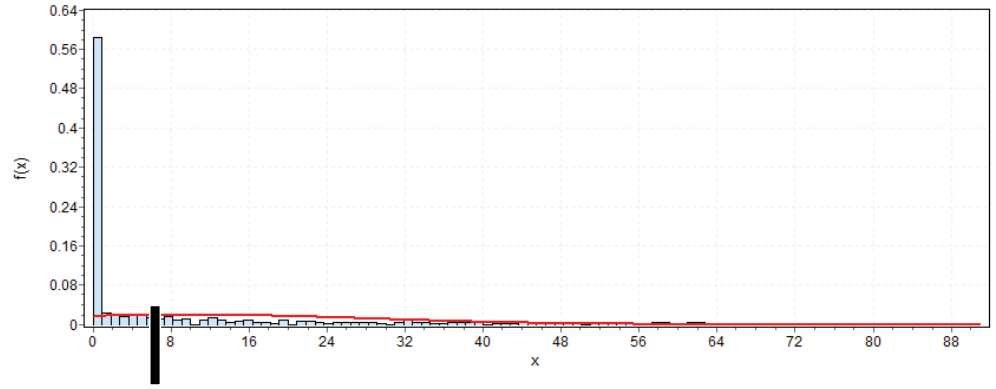
Slika 150. Raspodela podataka 7. atributa DP (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



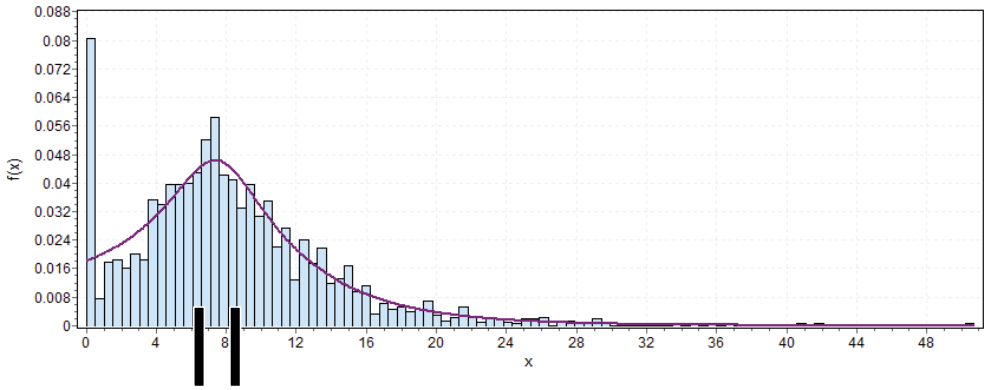
Slika 151. Raspodela podataka 8. atributa ASTV (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



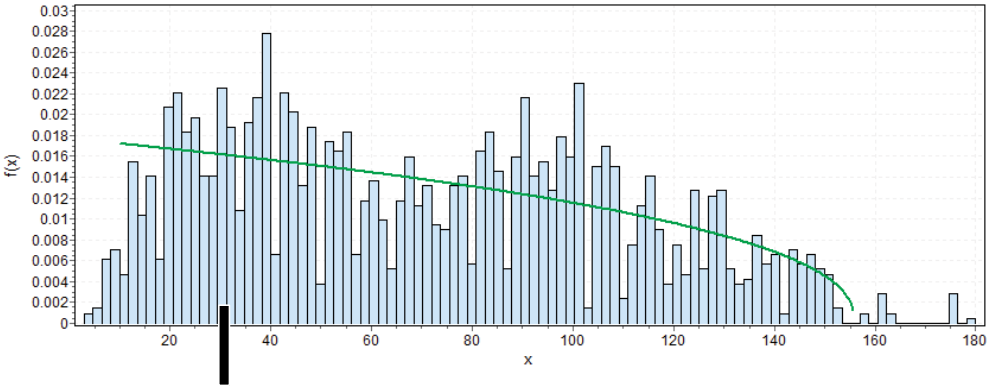
Slika 152. Raspodela podataka 9. atributa MSTV (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



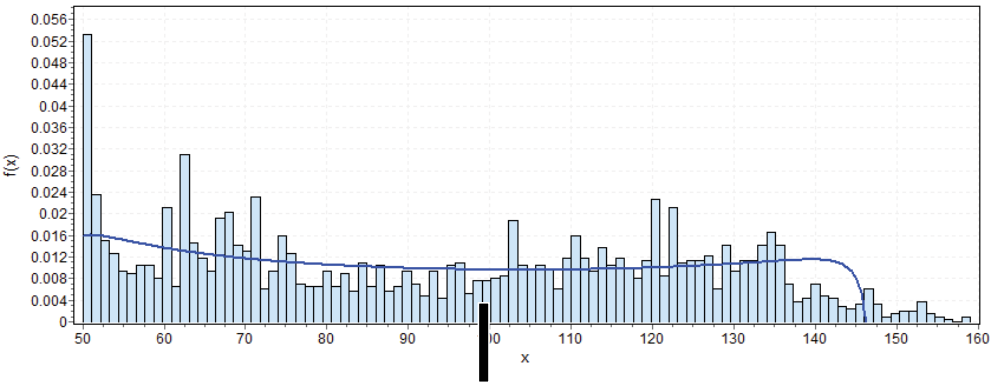
Slika 153. Raspodela podataka 10. atributa ALTV (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



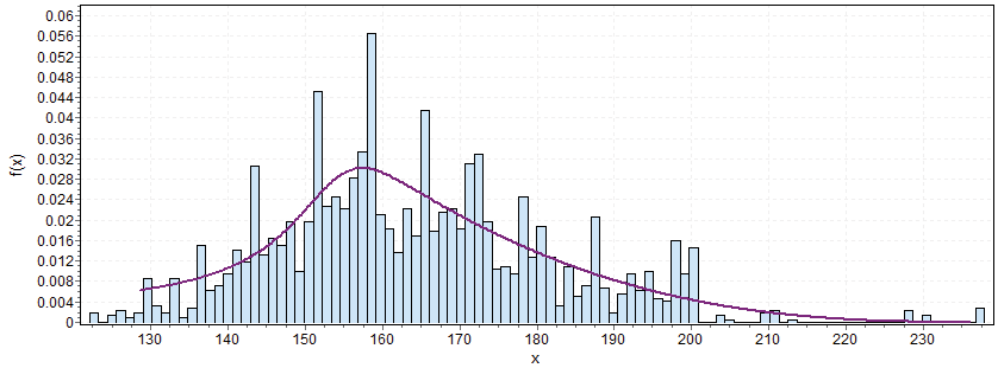
Slika 154. Raspodela podataka 11. atributa MLTV (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



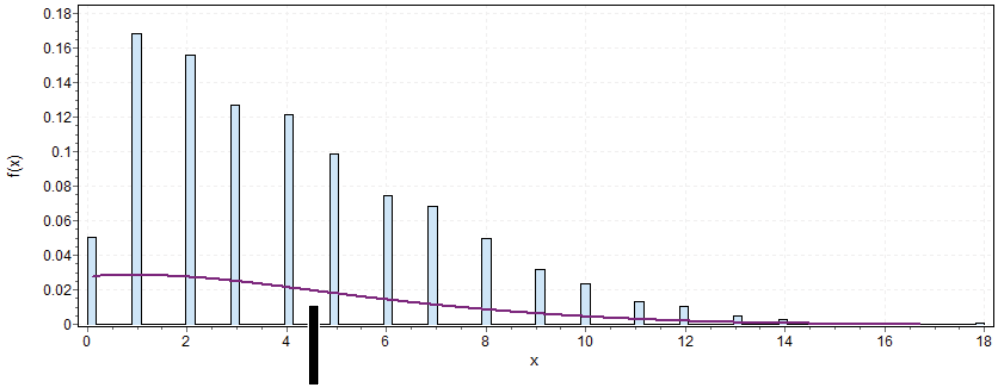
Slika 155. Raspodela podataka 12. atributa Width (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



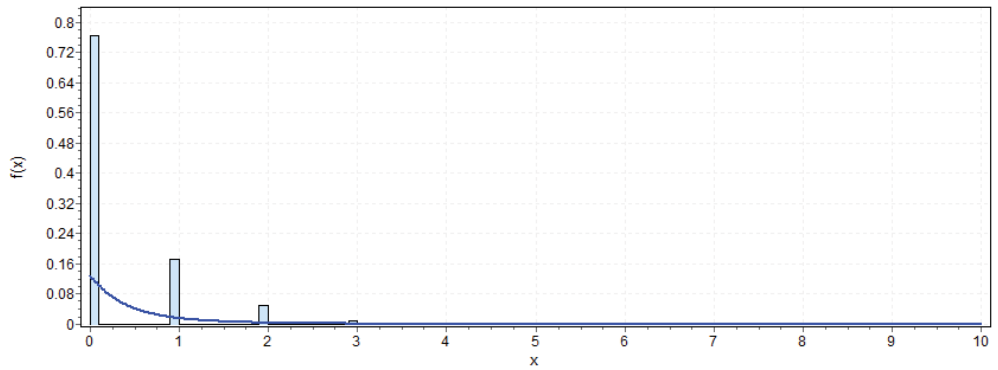
Slika 156. Raspodela podataka 13. atributa Min (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



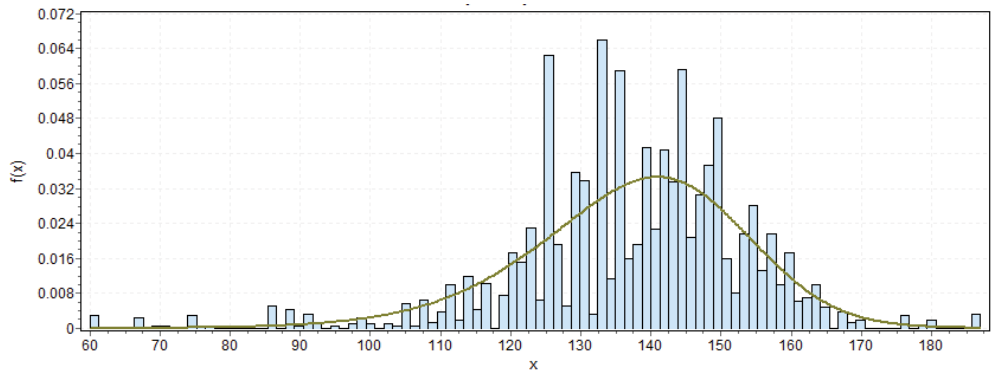
Slika 157. Raspodela podataka 14. atributa Max (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



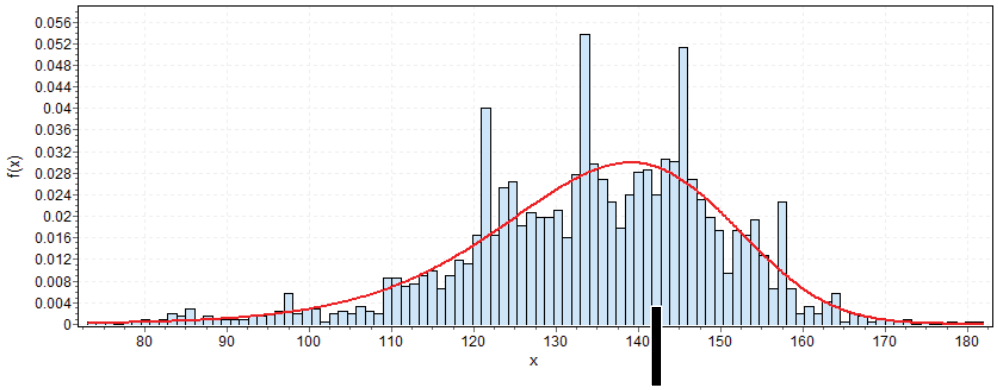
Slika 158. Raspodela podataka 15. atributa Nmax (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



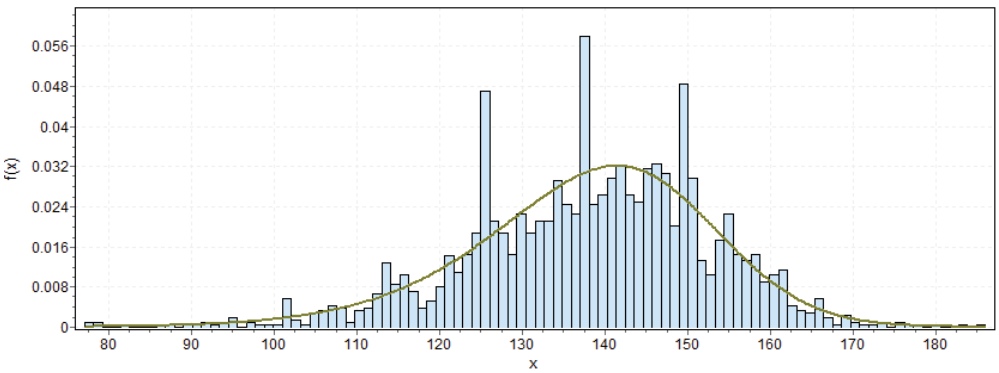
Slika 159. Raspodela podataka 16. atributa Nzeros (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



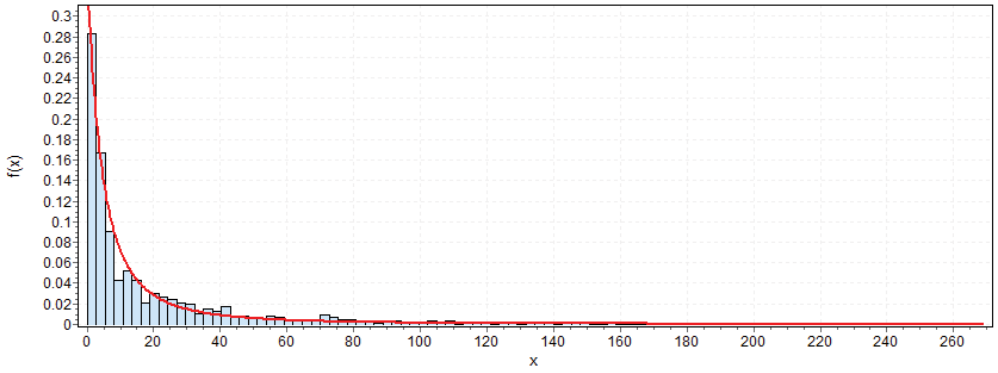
Slika 160. Raspodela podataka 17. atributa Mode (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



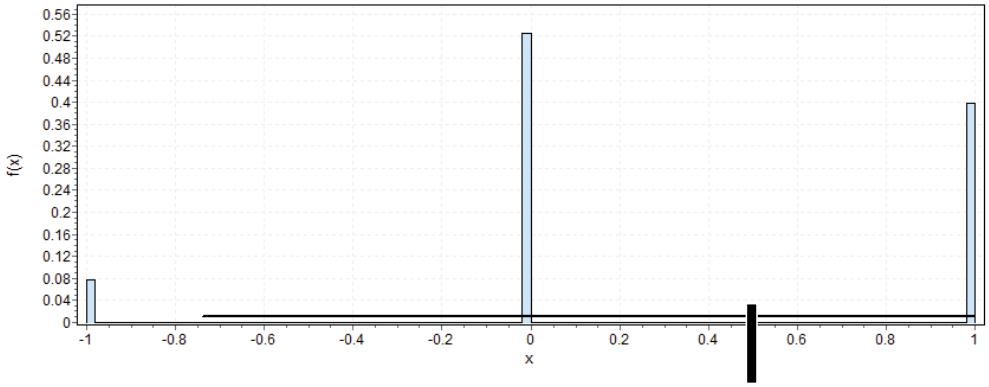
Slika 161. Raspodela podataka 18. atributa Mean (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



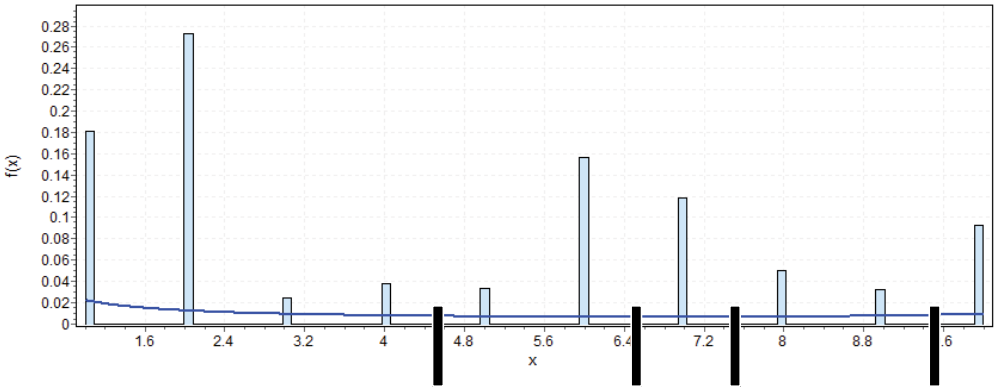
Slika 162. Raspodela podataka 19. atributa Median (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



Slika 163. Raspodela podataka 20. atributa Variance (baze Cardiotocography) nema tačke reza na osnovu algoritma maksimalne razberivosti



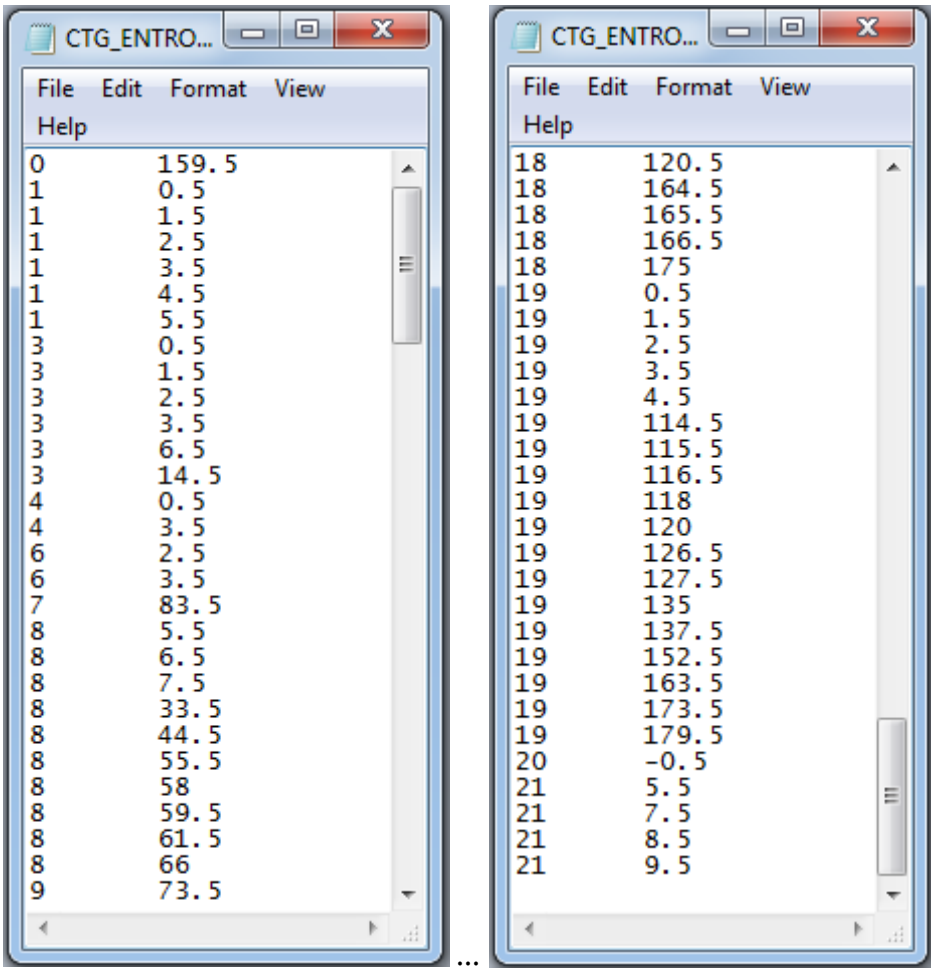
Slika 164. Raspodela podataka 21. atributa Tendency (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



Slika 165. Raspodela podataka 22. atributa CLASS (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

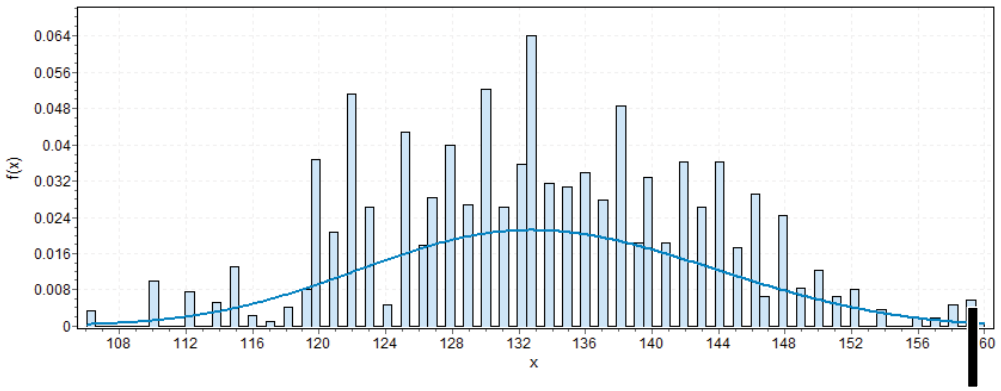
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Glass Identification algoritmom baziranim na entropiji, dobijene su tačke reza, čiji deo je prikazan na slici 166..

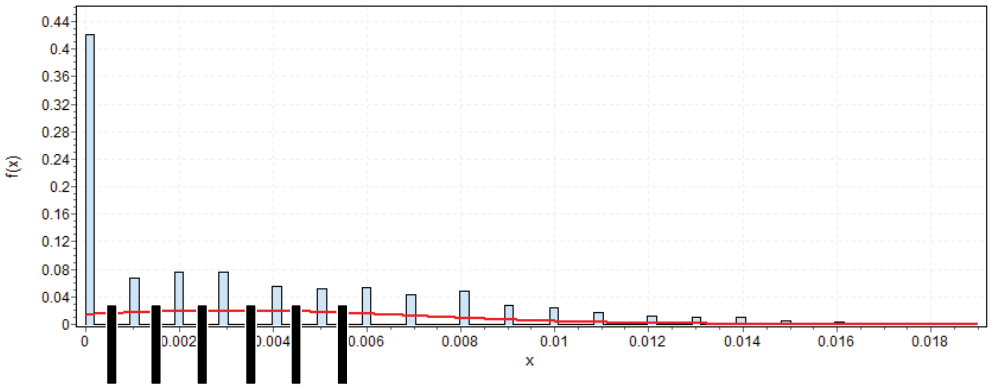


Slika 166. Deo tačaka reza baze Cardiotocography dobijene algoritmom baziranim na entropiji

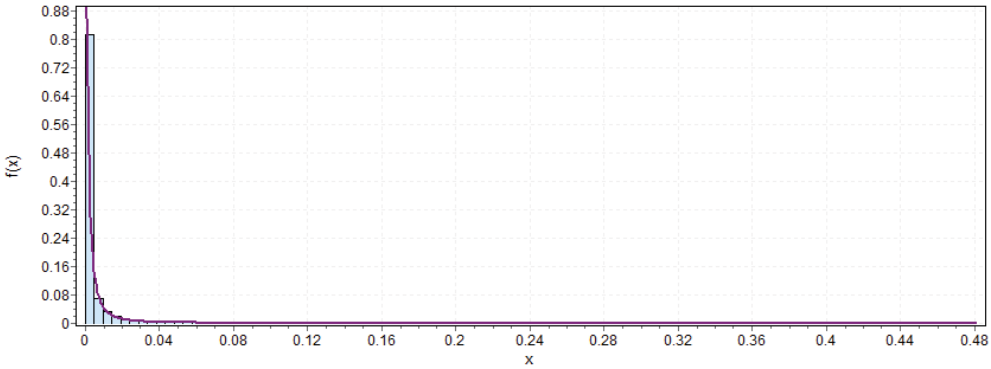
U odnosu na raspodelu podataka baze Cardiotocography koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 167 - 188).



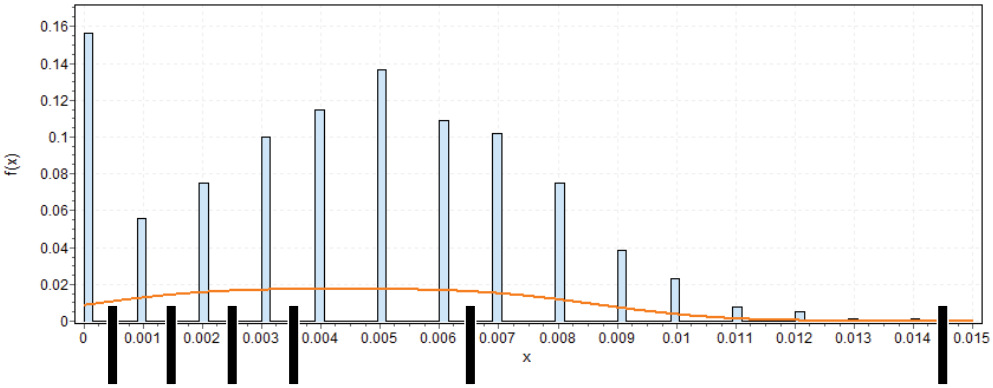
Slika 167. Raspodela podataka 1. atributa LB (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



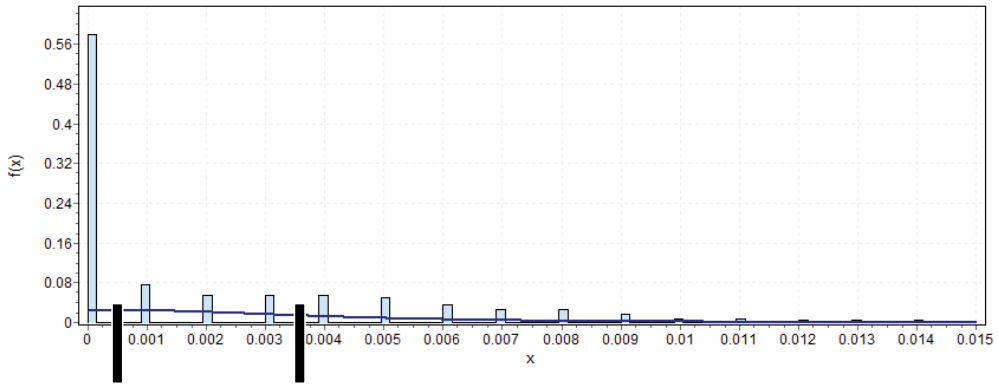
Slika 168. Raspodela podataka 2. atributa AC (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



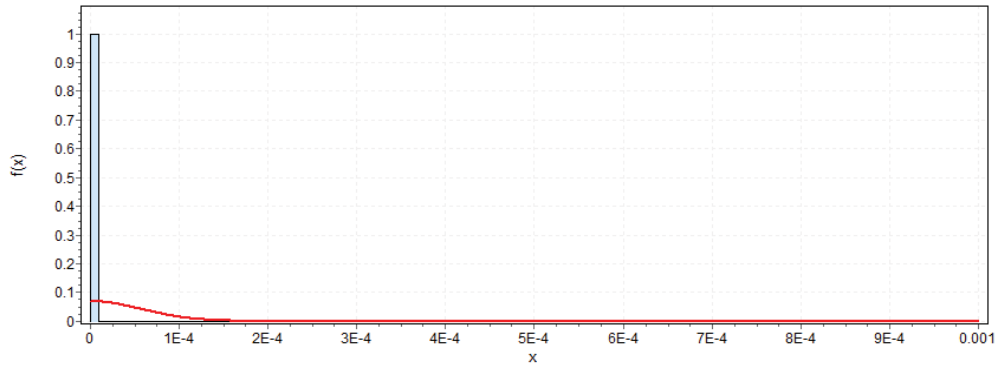
Slika 169. Raspodela podataka 3. atributa FM (baze Cardiotocography) nema tačaka reza na osnovu algoritma baziranog na entropiji



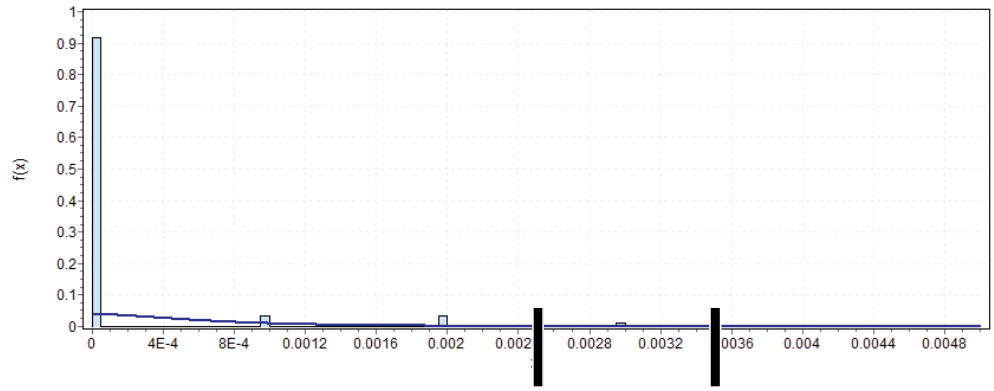
Slika 170. Raspodela podataka 4. atributa UC (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



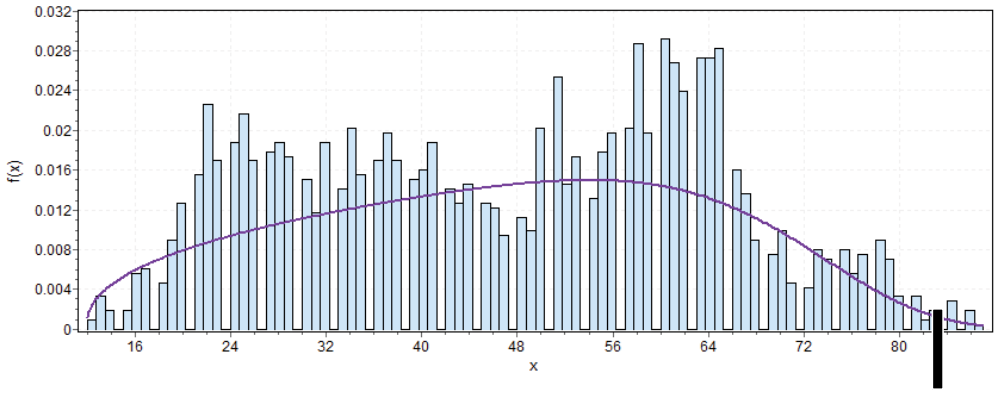
Slika 171. Raspodela podataka 5. atributa DL (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



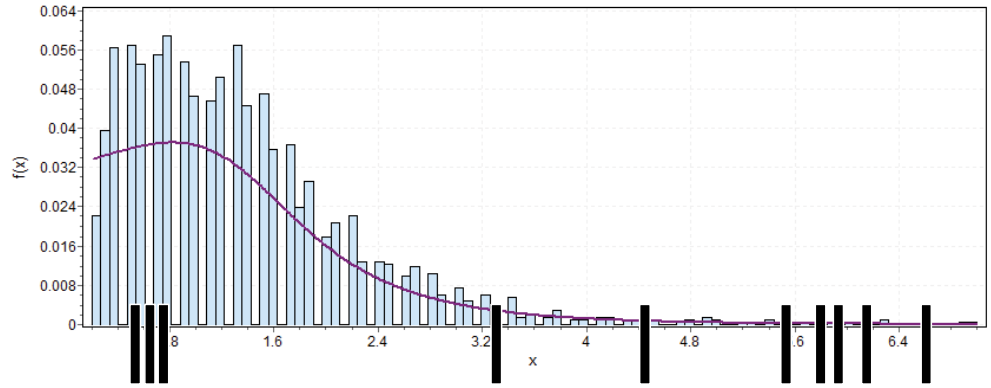
Slika 172. Raspodela podataka 6. atributa DS (baze Cardiotocography) nema tačaka reza na osnovu algoritma baziranog na entropiji



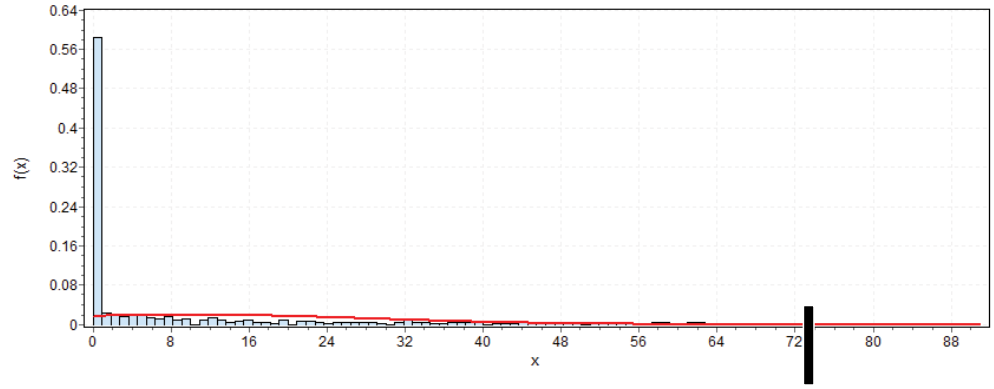
Slika 173. Raspodela podataka 7. atributa DP (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



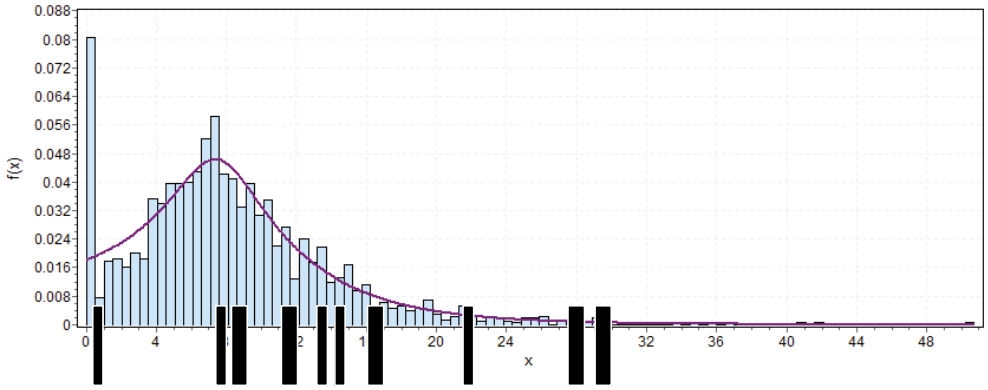
Slika 174. Raspodela podataka 8. atributa ASTV (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



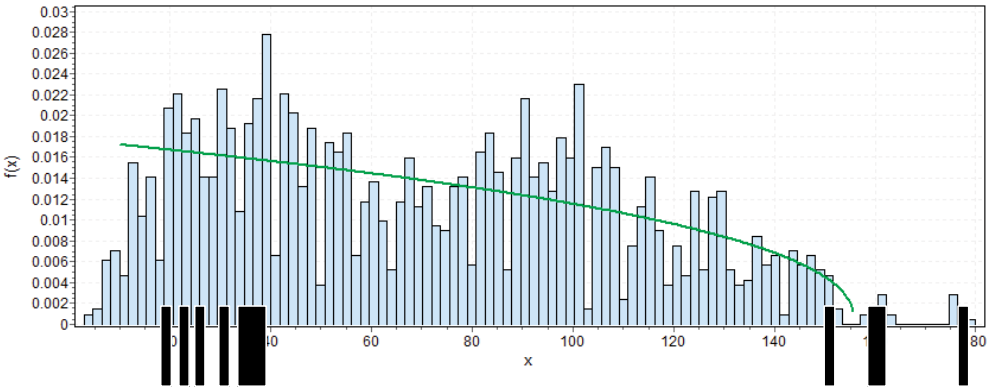
Slika 175. Raspodela podataka 9. atributa MSTV (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



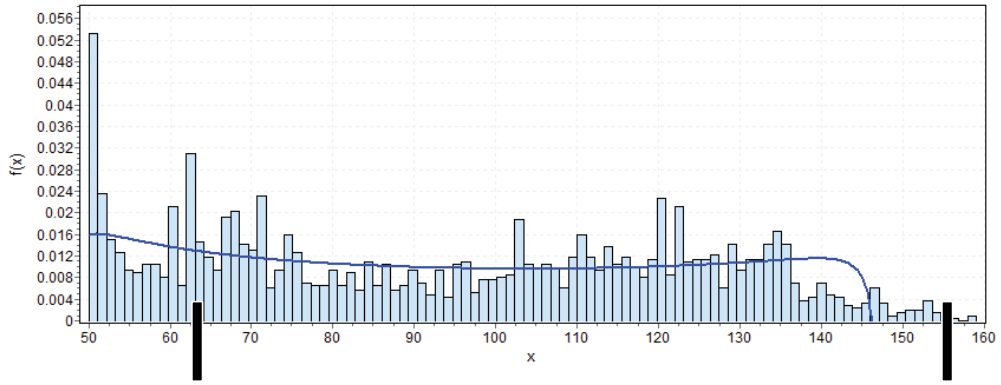
Slika 176. Raspodela podataka 10. atributa ALTV (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



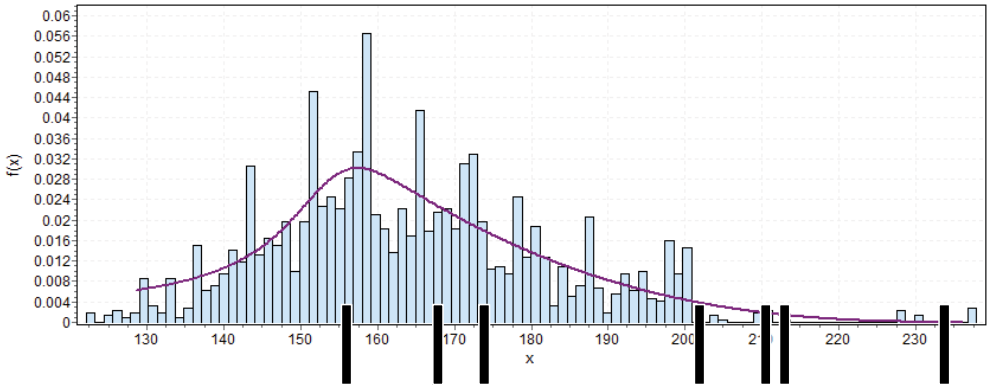
Slika 177. Raspodela podataka 11. atributa MLTV (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



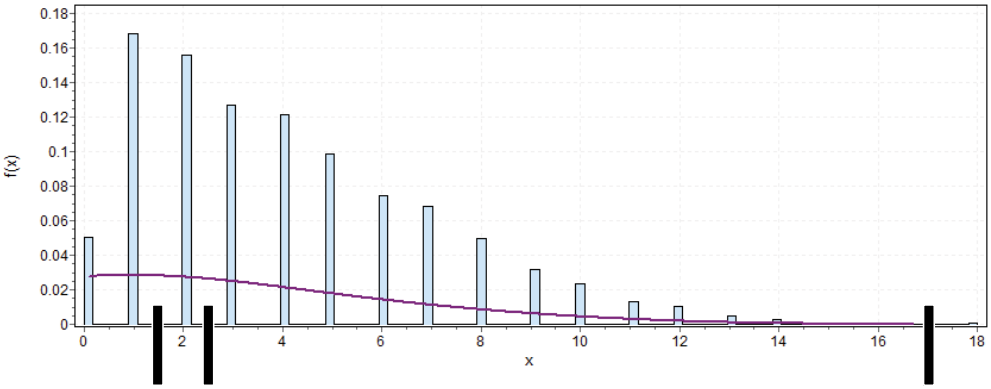
Slika 178. Raspodela podataka 12. atributa Width (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



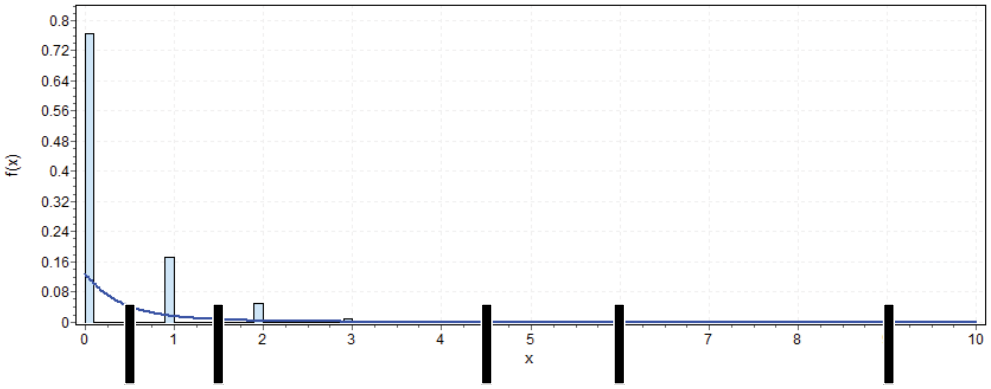
Slika 179. Raspodela podataka 13. atributa Min (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



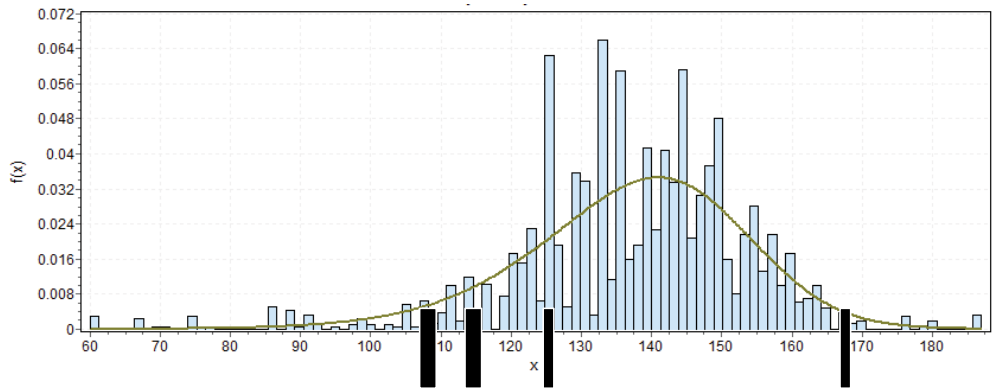
Slika 180. Raspodela podataka 14. atributa Max (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



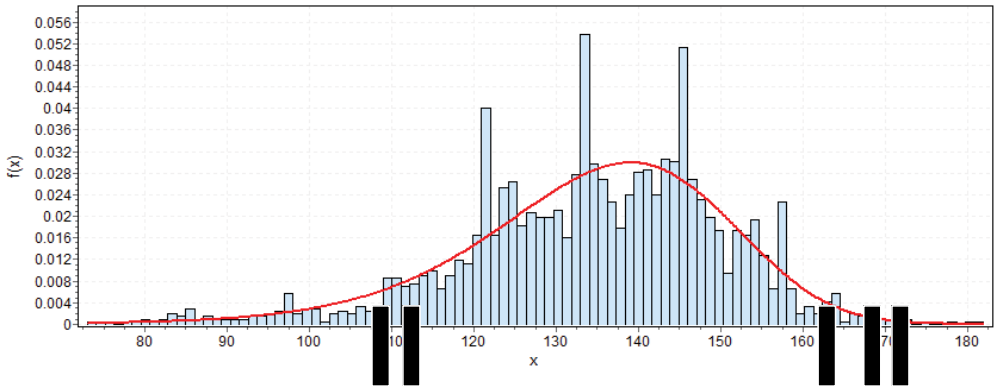
Slika 181. Raspodela podataka 15. atributa Nmax (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



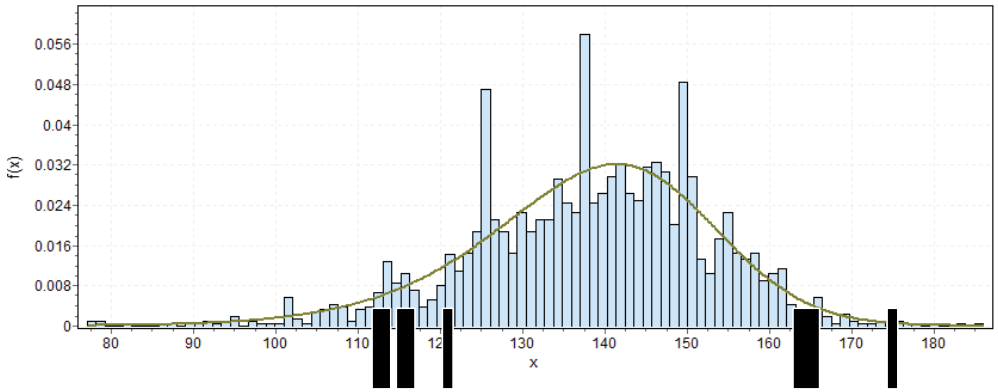
Slika 182. Raspodela podataka 16. atributa Nzeros (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



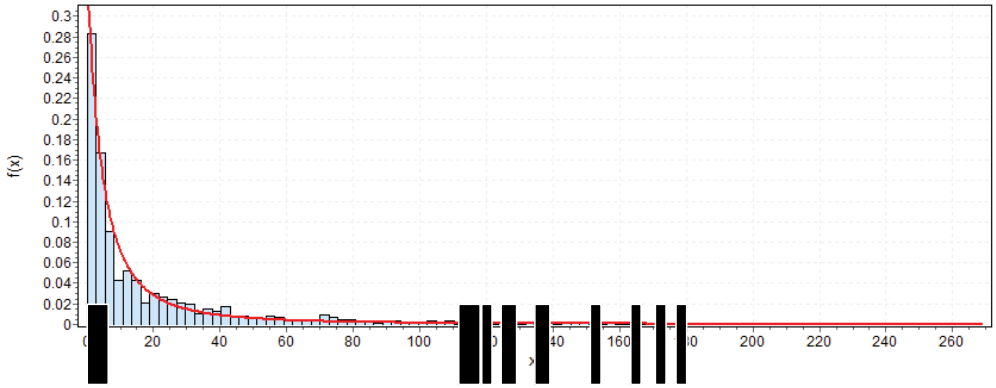
Slika 183. Raspodela podataka 17. atributa Mode (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



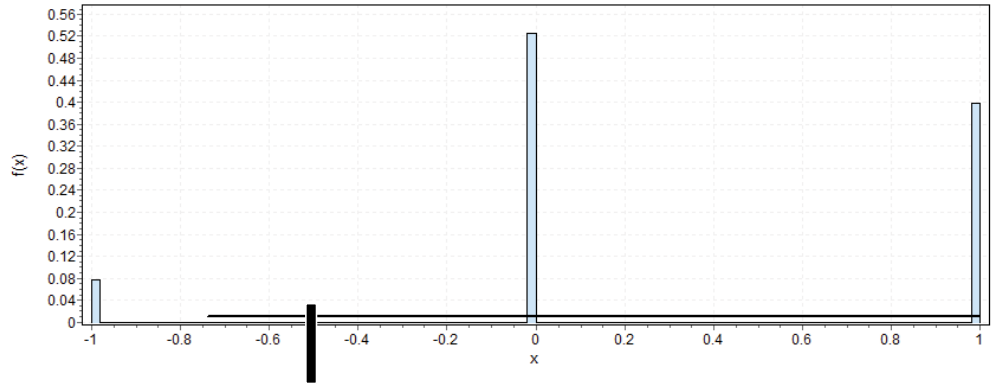
Slika 184. Raspodela podataka 18. atributa Mean (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



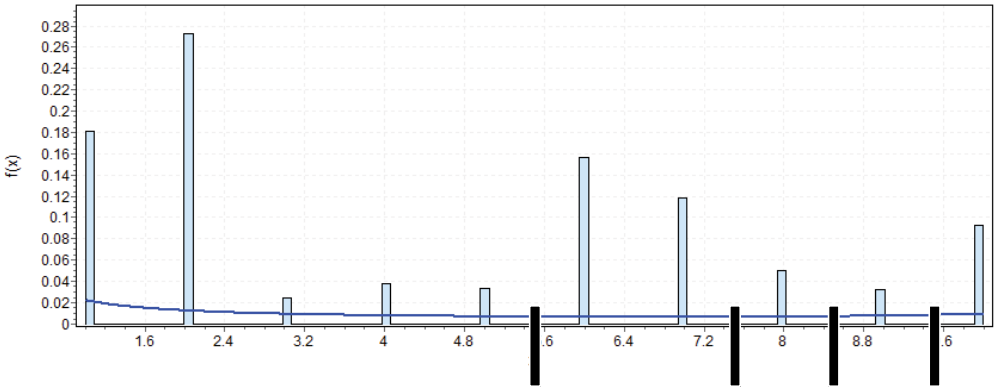
Slika 185. Raspodela podataka 19. atributa Median (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 186. Raspodela podataka 20. atributa Variance (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 187. Raspodela podataka 21. atributa Tendency (baze Cardiotocography) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



Slika 188. Raspodela podataka 22. atributa CLASS (baze Cardiotocography) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji

8. DODACI ZA ANALIZU BAZE STATLOG (AUSTRALIAN CREDIT APPROVAL)

Bazu Statlog (Australian Credit Approval) je donirao J.R. QUINLAN, Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney, Australia. Ova baza je interesantna zbog dobrog miksa atributa – kontunualni, nominalni sa malim brojem vrednosti i nominalni sa velikim brojemvrednosti [Statlog, 2015].

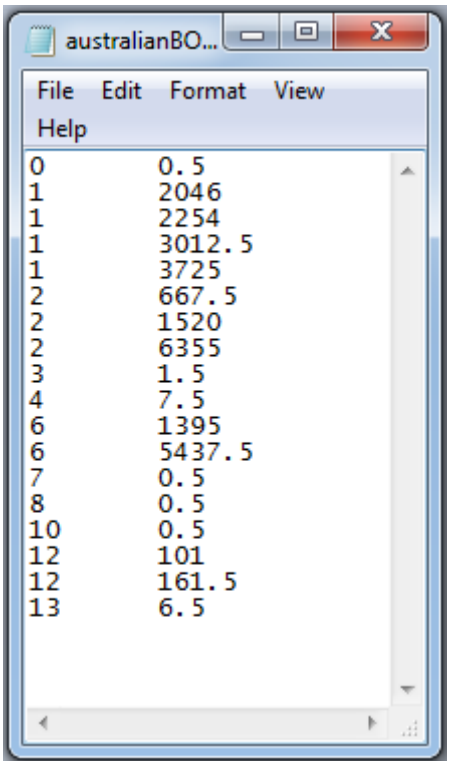
Informacije o atributima (preuzeto iz [Statlog, 2015])

A1: 0,1 CATEGORICAL (formerly: a,b)
A2: continuous.
A3: continuous.
A4: 1,2,3 CATEGORICAL (formerly: p,g,gg)
A5: 1, 2,3,4,5, 6,7,8,9,10,11,12,13,14 CATEGORICAL (formerly: ff,d,i,k,j,aa,m,c,w, e, q, r,cc, x)
A6: 1, 2,3, 4,5,6,7,8,9 CATEGORICAL (formerly: ff,dd,j,bb,v,n,o,h,z)
A7: continuous.
A8: 1, 0 CATEGORICAL (formerly: t, f)
A9: 1, 0 CATEGORICAL (formerly: t, f)
A10: continuous.
A11: 1, 0 CATEGORICAL (formerly t, f)
A12: 1, 2, 3 CATEGORICAL (formerly: s, g, p)
A13: continuous.
A14: continuous.
A15: 1,2 class attribute (formerly: +,-)

Broj instanci je 690. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

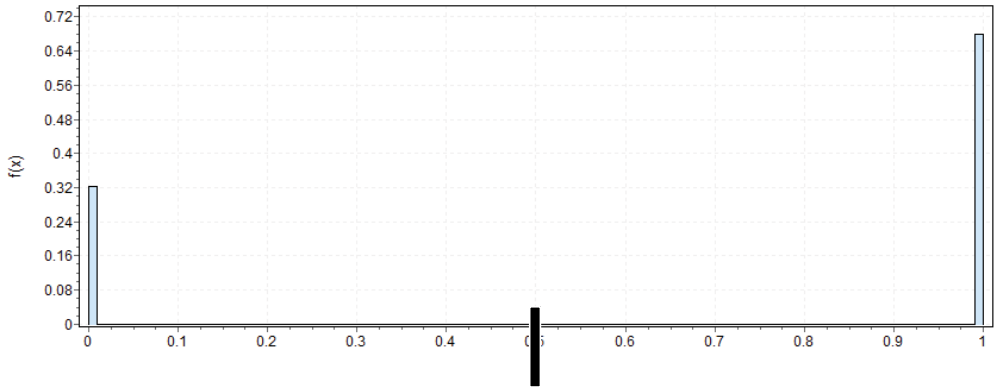
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Statlog (Australian Credit Approval) algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 189.

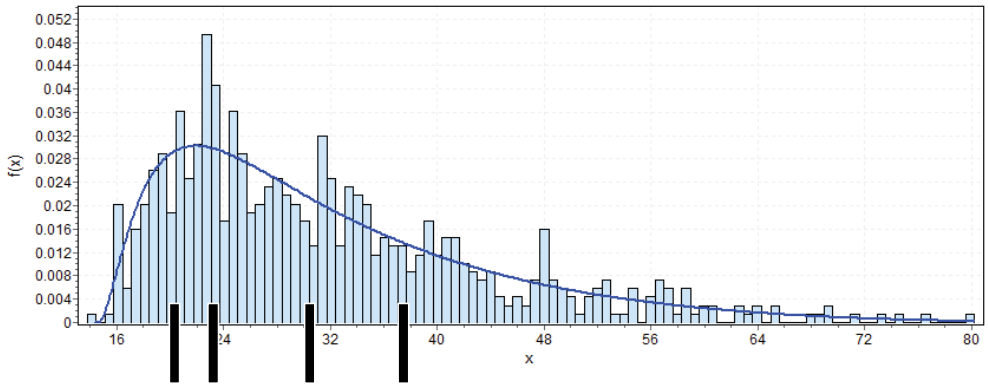


Slika 189. Tačke reza baze Statlog (Australian Credit Approval) dobijene algoritmom maksimalne razberivosti

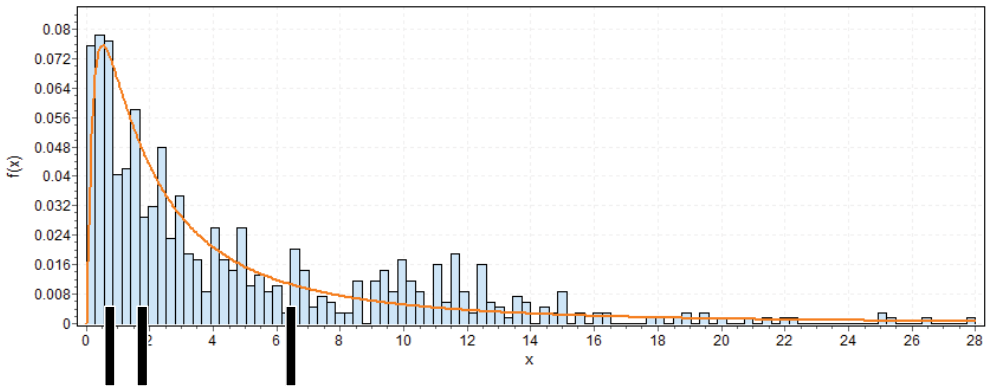
U odnosu na raspodelu podataka baze Statlog (Australian Credit Approval) koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 190 - 203).



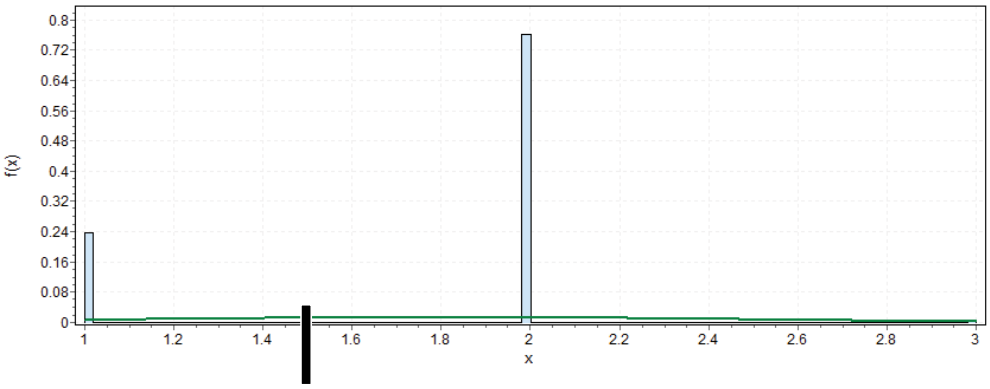
Slika 190. Raspodela podataka 1. atributa A1 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



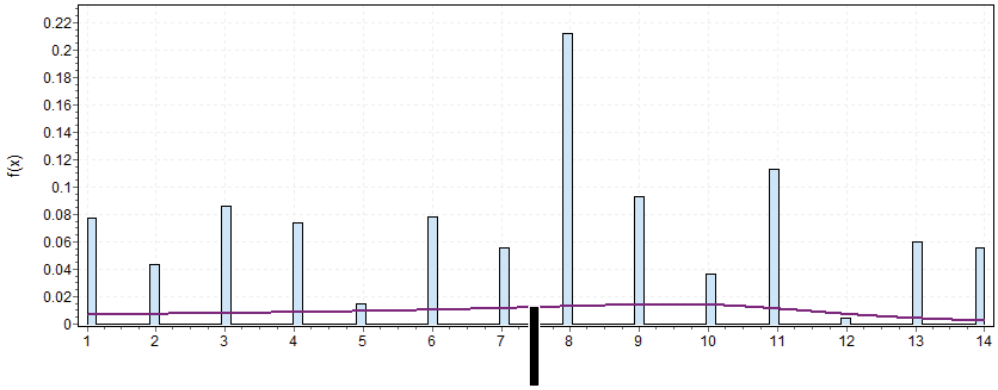
Slika 191. Raspodela podataka 2. atributa A2 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



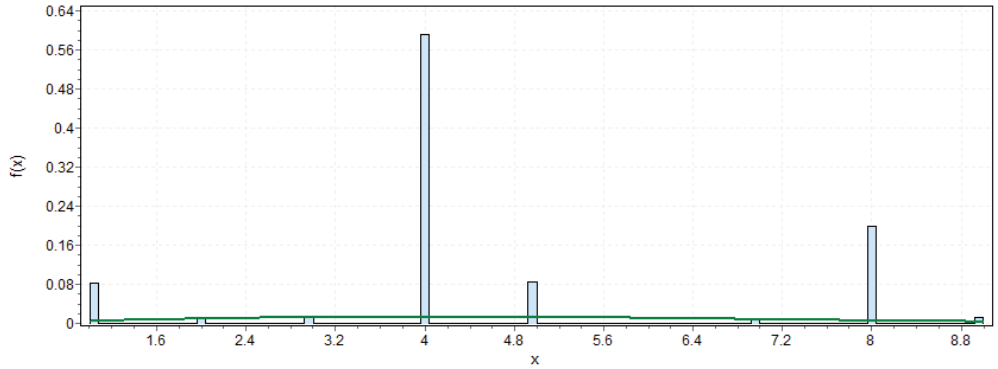
Slika 192. Raspodela podataka 3. atributa A3 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



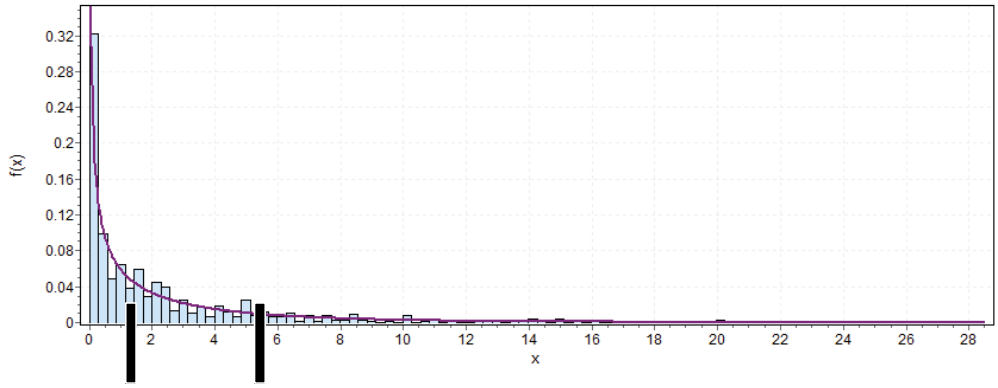
Slika 193. Raspodela podataka 4. atributa A4 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



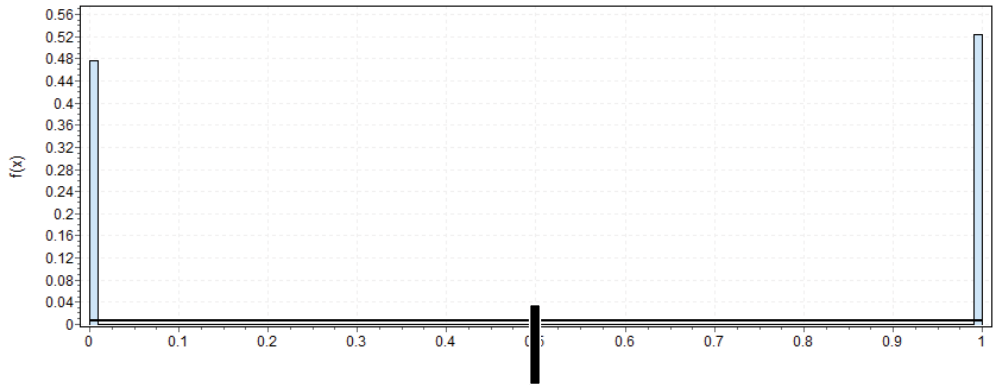
Slika 194. Raspodela podataka 5. atributa A5 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



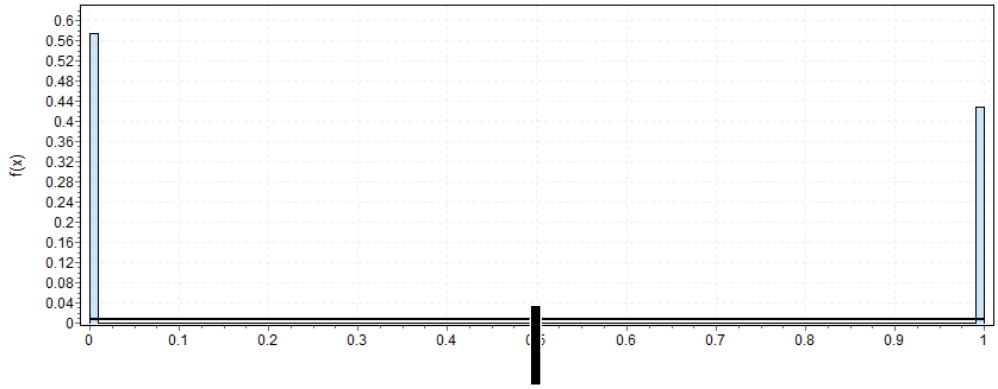
Slika 195. Raspodela podataka 6. atributa A6 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma maksimalne razberivosti



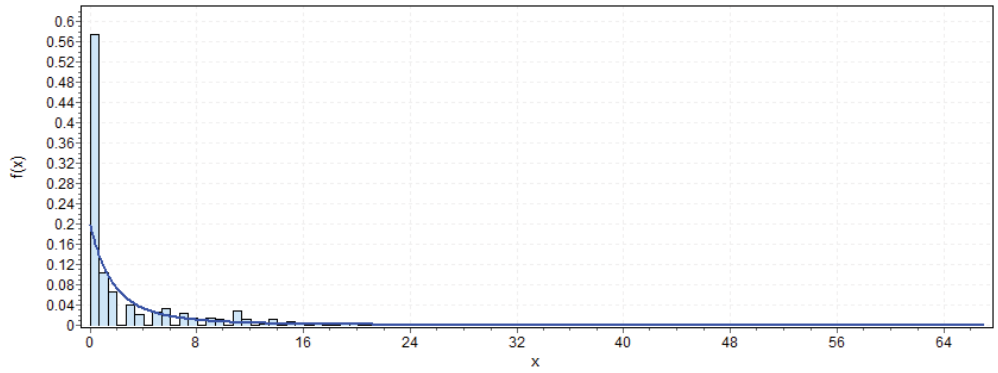
Slika 196. Raspodela podataka 7. atributa A7 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



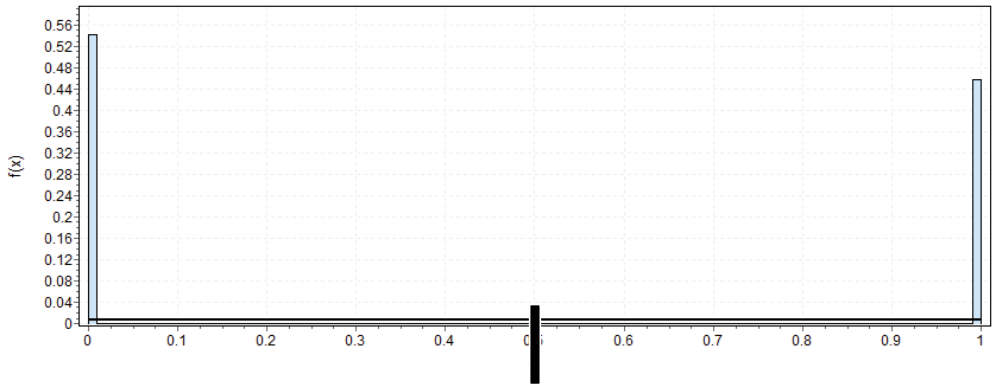
Slika 197. Raspodela podataka 8. atributa A8 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



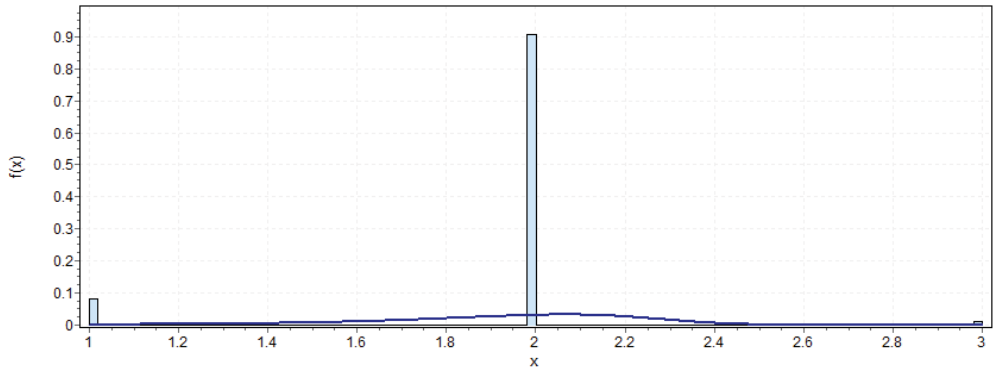
Slika 198. Raspodela podataka 9. atributa A9 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



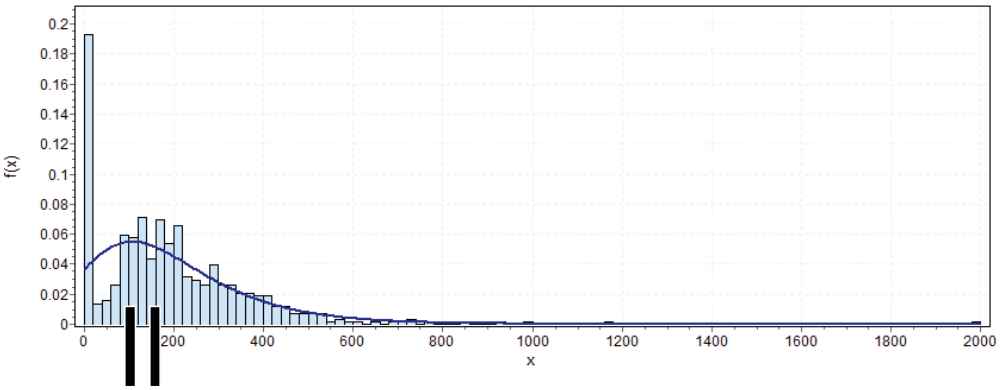
Slika 199. Raspodela podataka 10. atributa A10 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma maksimalne razberivosti



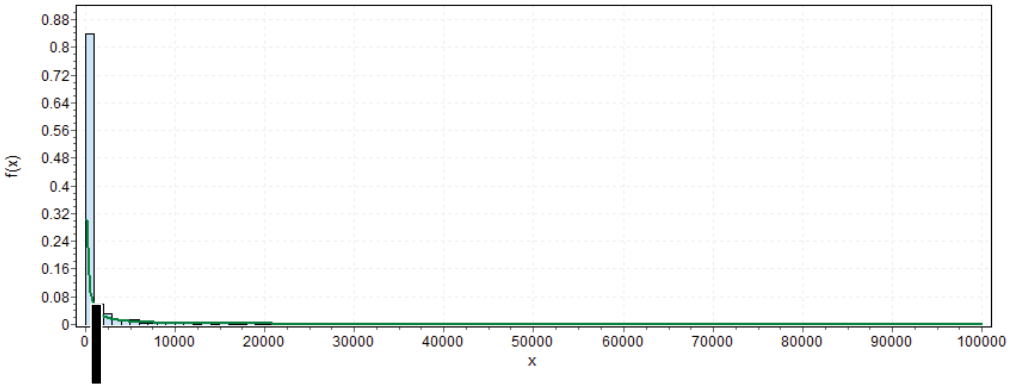
Slika 200. Raspodela podataka 11. atributa A11 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti



Slika 201. Raspodela podataka 12. atributa A12 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma maksimalne razberivosti



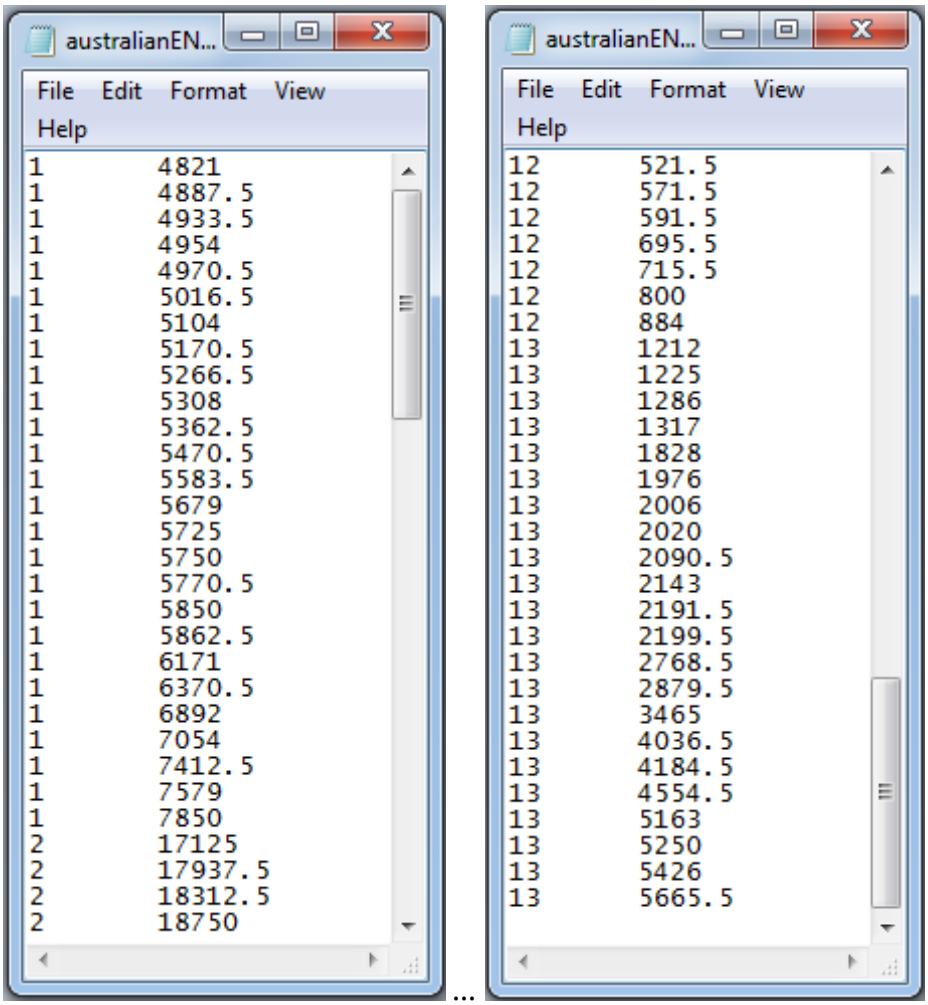
Slika 202. Raspodela podataka 13. atributa A13 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 203. Raspodela podataka 14. atributa A14 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma maksimalne razberivosti

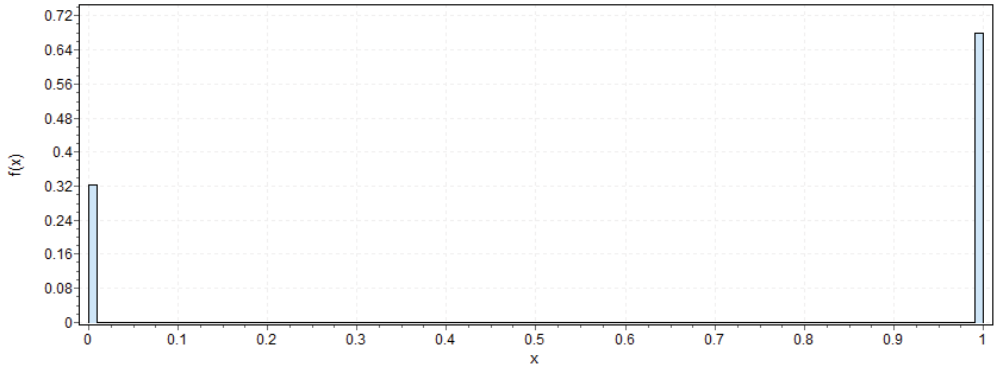
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Statlog (Australian Credit Approval) algoritmom baziranim na entropiji, dobijene su tačke reza čiji deo je prikazan na slici 204.

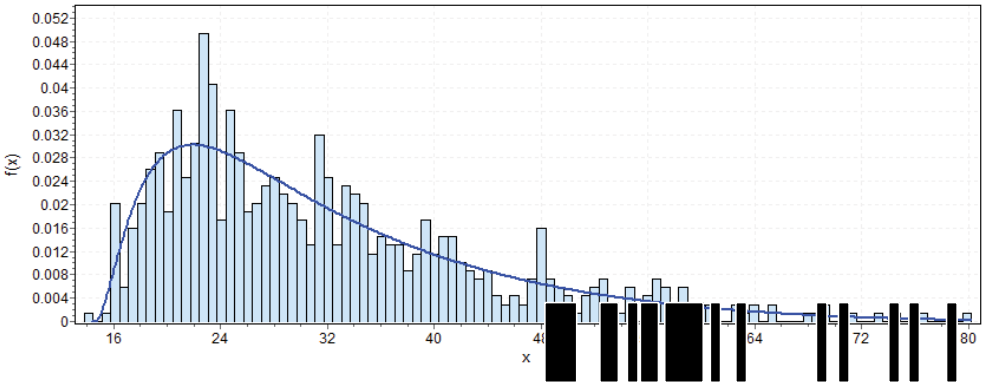


Slika 204. Deo tačska reza baze Statlog (Australian Credit Approval) dobijenih algoritmom baziranim na entropiji

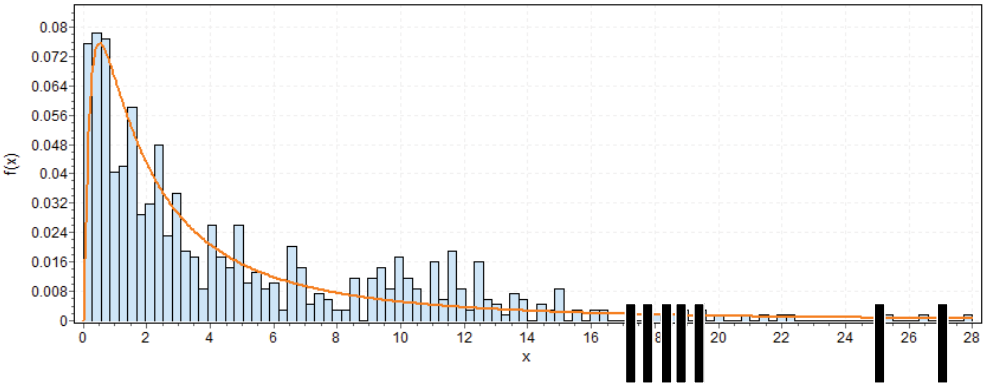
U odnosu na raspodelu podataka baze Statlog (Australian Credit Approval) koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 205 - 218).



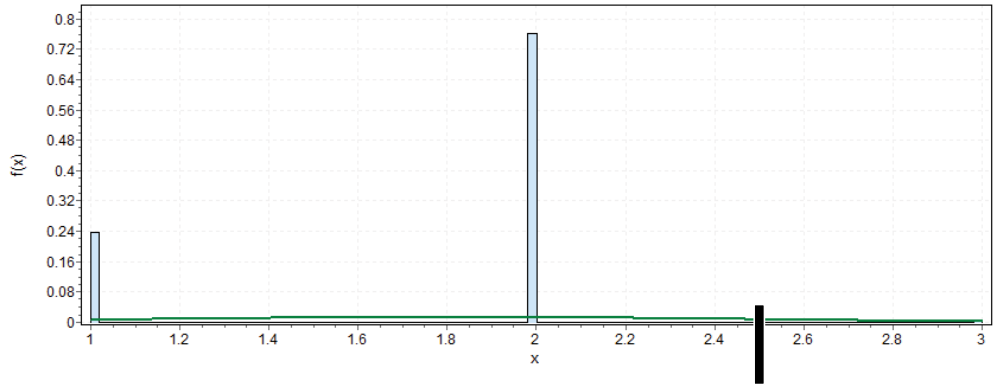
Slika 205. Raspodela podataka 1. atributa A1 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji na drugi način



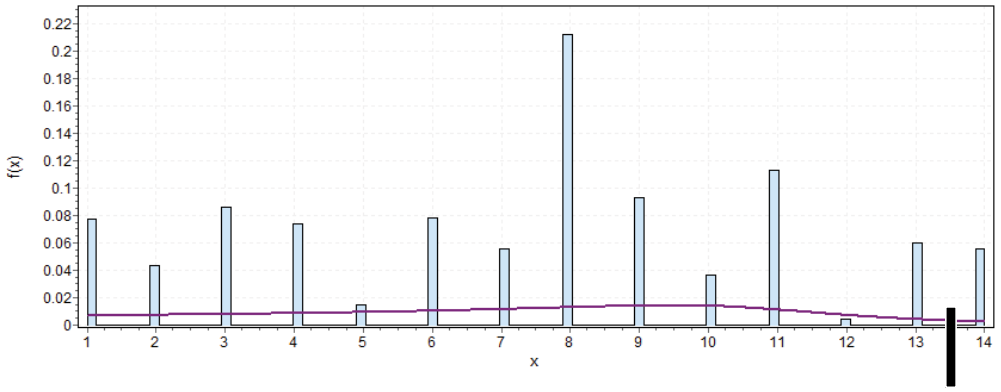
Slika 206. Raspodela podataka 2. atributa A2 (baze Statlog (Australian Credit Approval)) sa 26 tačaka reza dobijenim na osnovu algoritma baziranim na entropiji



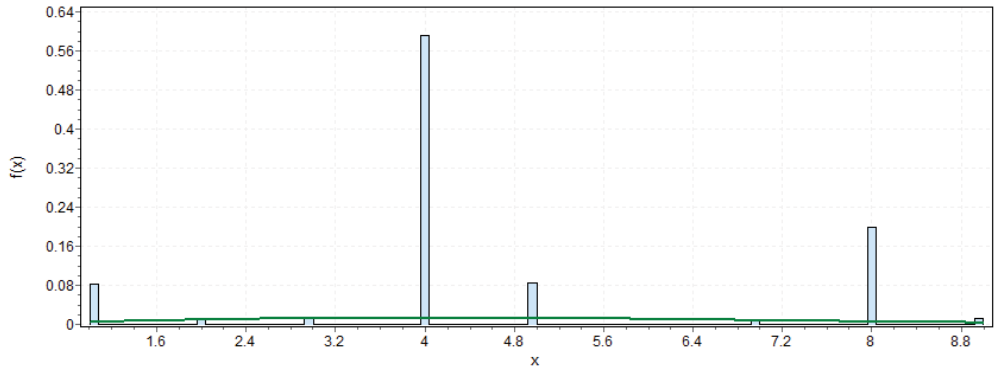
Slika 207. Raspodela podataka 3. atributa A3 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma baziranim na entropiji



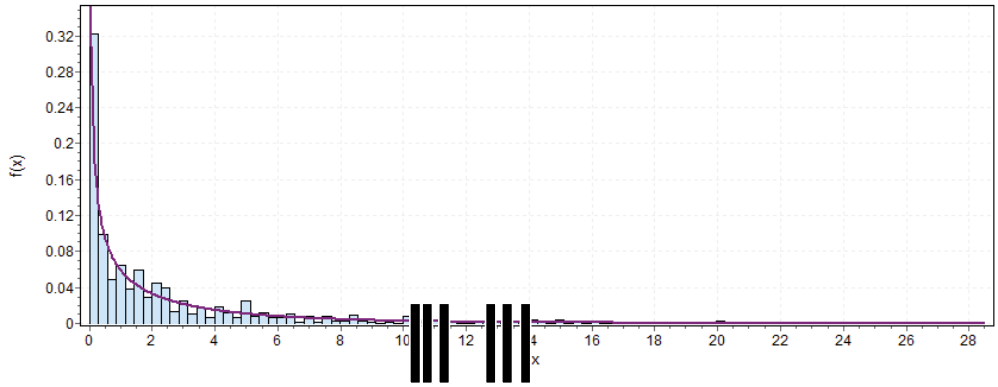
Slika 208. Raspodela podataka 4. atributa A4 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma baziranim na entropiji



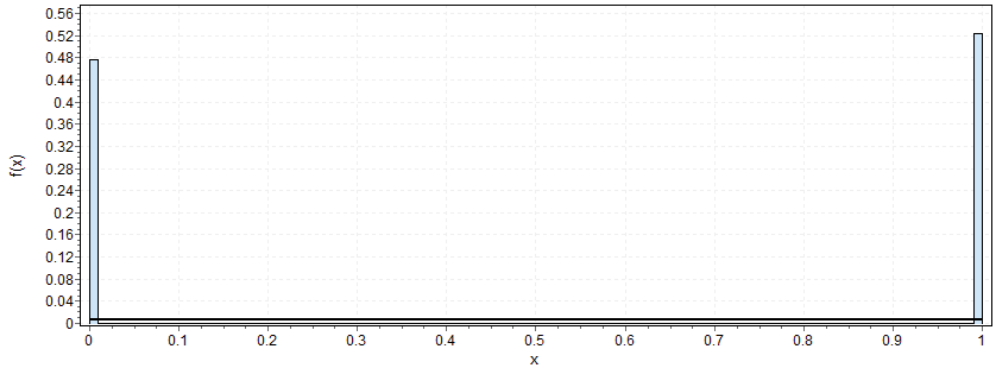
Slika 209. Raspodela podataka 5. atributa A5 (baze Statlog (Australian Credit Approval)) sa tačkom reza dobijenom na osnovu algoritma baziranim na entropiji



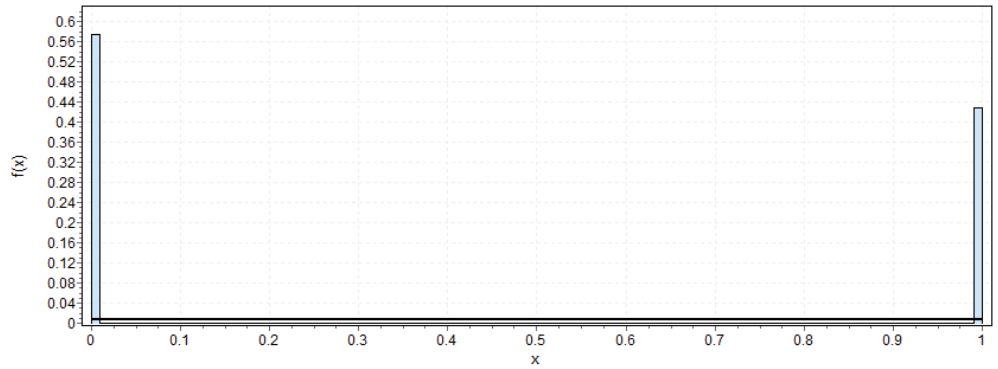
Slika 210. Raspodela podataka 6. atributa A6 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji



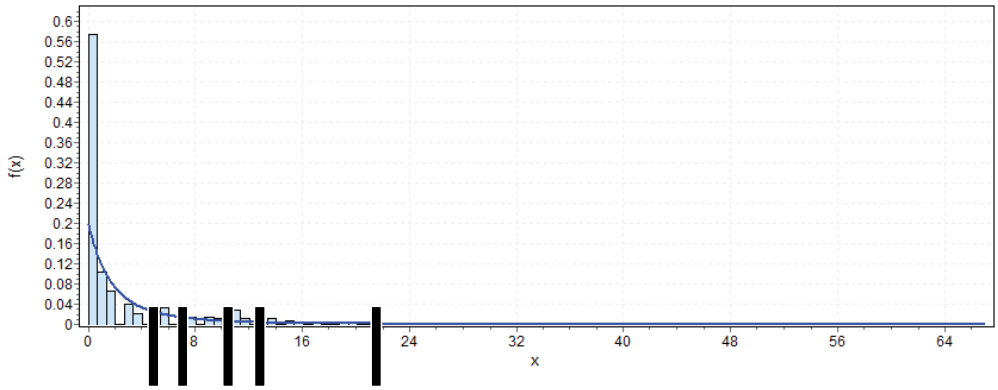
Slika 211. Raspodela podataka 7. atributa A7 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma baziranim na entropiji



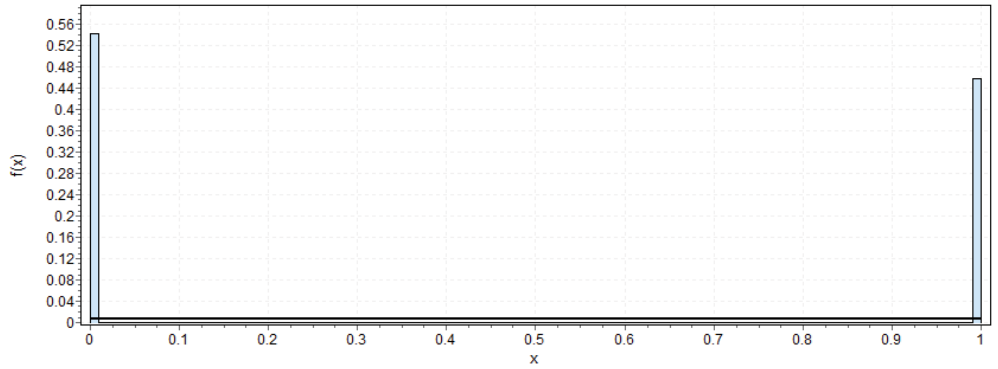
Slika 212. Raspodela podataka 8. atributa A8 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji



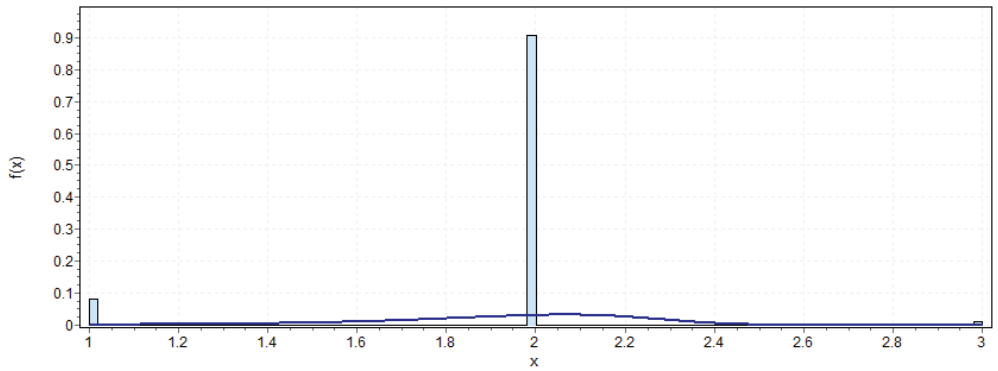
Slika 213. Raspodela podataka 9. atributa A9 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji



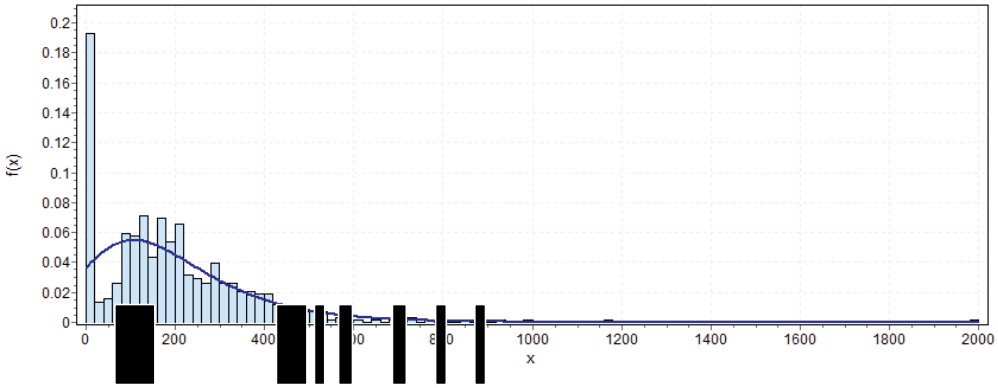
Slika 214. Raspodela podataka 10. atributa A10 (baze Statlog (Australian Credit Approval)) sa tačkama reza dobijenim na osnovu algoritma baziranim na entropiji



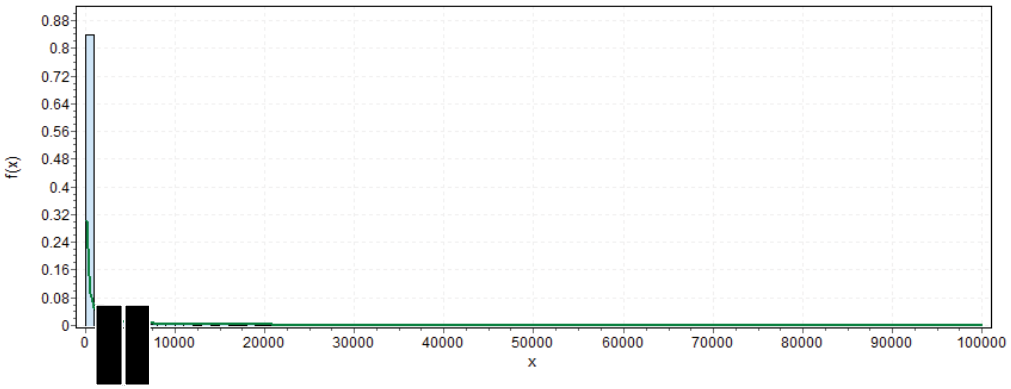
Slika 215. Raspodela podataka 11. atributa A11 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji



Slika 216. Raspodela podataka 12. atributa A12 (baze Statlog (Australian Credit Approval)) bez tačke reza na osnovu algoritma baziranim na entropiji



Slika 217. Raspodela podataka 13. atributa A13 (baze Statlog (Australian Credit Approval)) sa 23 tačke reza dobijenim na osnovu algoritma baziranim na entropiji



Slika 218. Raspodela podataka 14. atributa A14 (baze Statlog (Australian Credit Approval)) sa 22 tačke reza dobijenih na osnovu algoritma baziranim na entropiji

9. DODACI ZA ANALIZU BAZE HABERMAN'S SURVIVAL DATA SET

Baza sadrži podatke istraživanja koje je sprovedeno između 1958. i 1970. godine na Univerzitetu Chicago's Billings Hospital i odnosi se na preživljanje pacijentkinja koje su bile podvrgnute operaciji raka dojke. Donor je Tjen-Sien Lim.

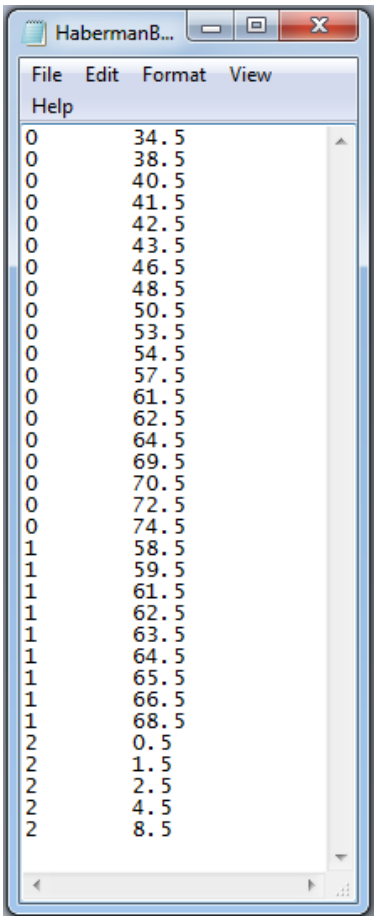
Informacije o atributima (preuzeto iz [Haberman, 1999]):

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
-- 1 = the patient survived 5 years or longer
-- 2 = the patient died within 5 year

Broj instanci je 306. U sistemu Rosetta urađena je diskretizacija ove baze i dobijeni su sledeći rezultati:

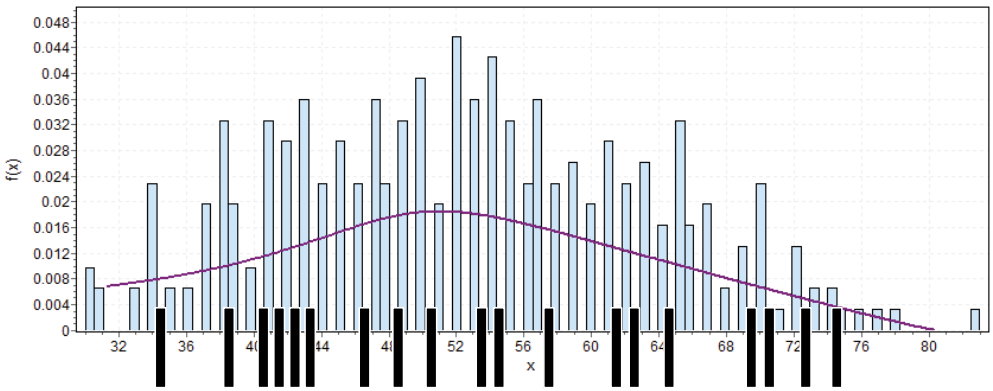
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Haberman's Survival Data Set algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 219.

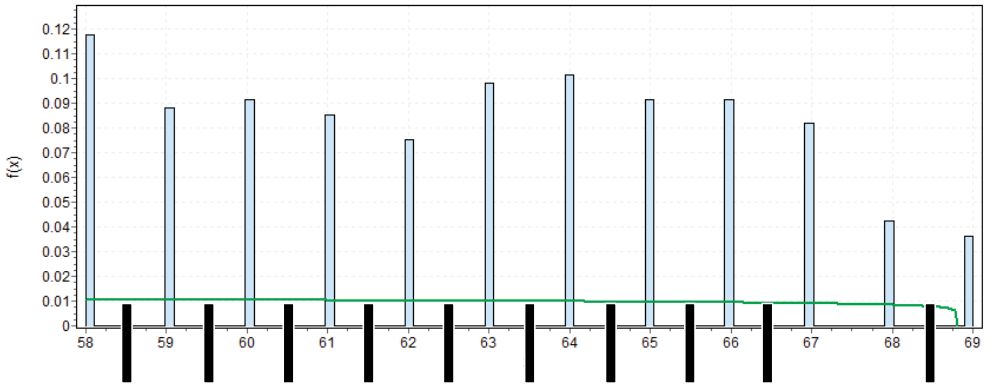


Slika 219. Tačke reza baze Haberman's Survival Data Set dobijene algoritmom maksimalne razberivosti

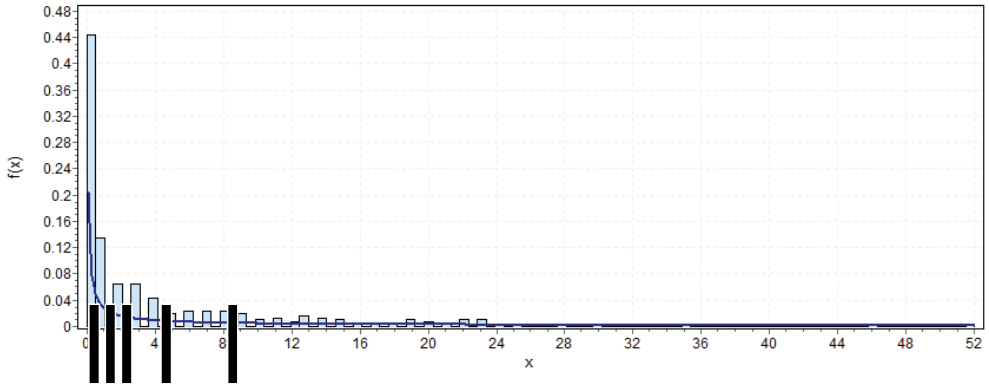
U odnosu na raspodelu podataka baze Haberman's Survival Data Set koja je analizirana softverom EasyFit, ove tačke reza su prikazane na ordinatama širokim crnim vertikanim linijama (slike 220 - 222).



Slika 220. Raspodela podataka 1. atributa Age of patient at time of operation (baze Haberman's Survival Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 221. Raspodela podataka 2. atributa Patient's year of operation (baze Haberman's Survival Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

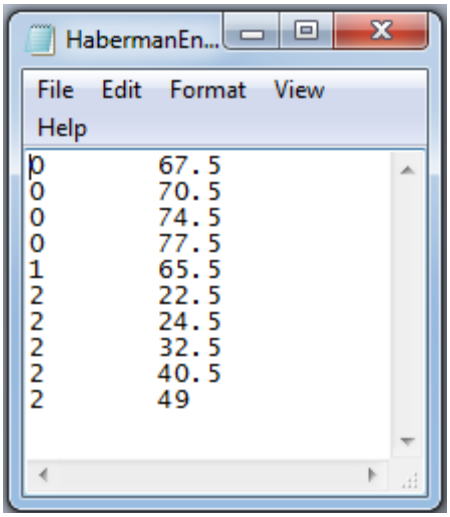


Slika 222. Raspodela podataka 3. atributa Number of positive axillary nodes detected (baze Haberman's Survival Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

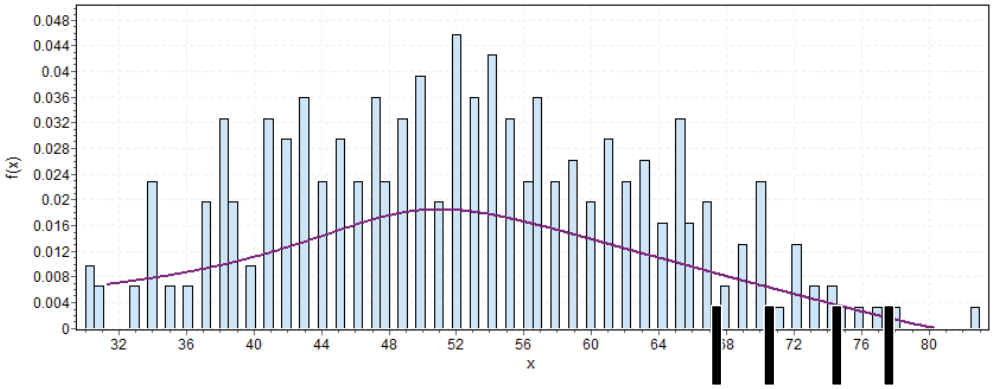
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Haberman's Survival Data Set algoritmom baziranim na entropiji, dobijene su tačke reza, čiji deo je prikazan na slici 223.

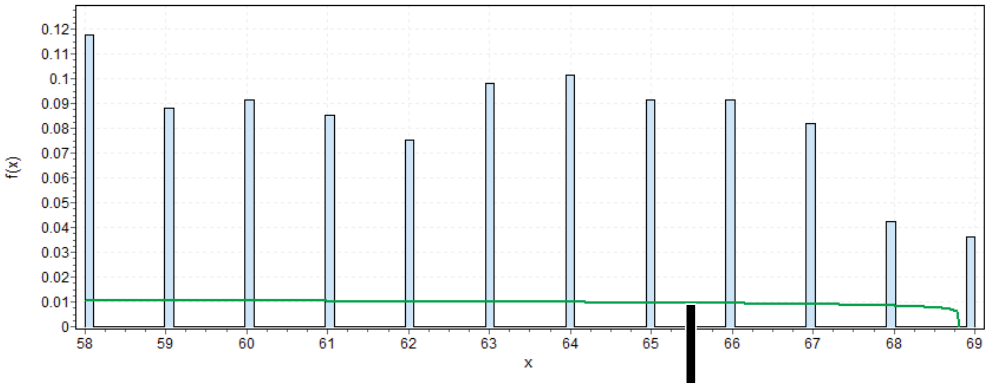
U odnosu na raspodelu podataka baze Glass Identification koja je analizirana softverom EasyFit, ove tačke reza prikazane su na ordinatama širokim crnim vertikanim linijama (slike 224 - 226).



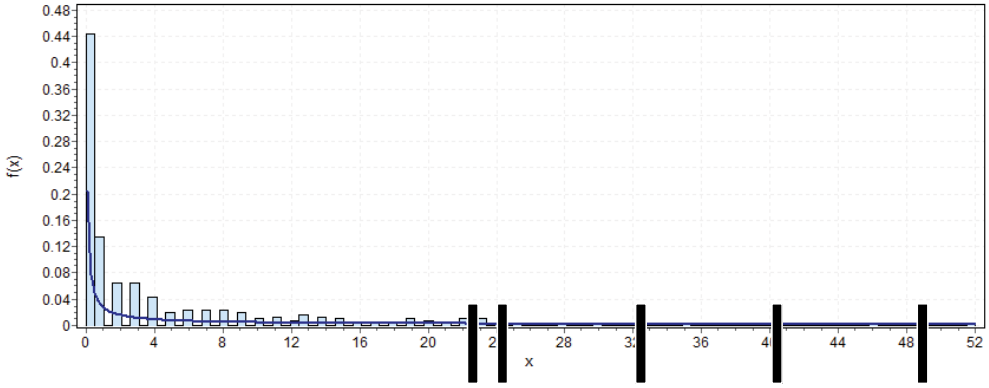
Slika 223. Tačke reza baze Haberman's Survival Data Set dobijene algoritmom baziranim na entropiji



Slika 224. Raspodela podataka 1. atributa Age of patient at time of operation (baze Haberman's Survival Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji



Slika 225. Raspodela podataka 2. atributa Patient's year of operation (baze Haberman's Survival Data Set) sa tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



Slika 226. Raspodela podataka 3. atributa Number of positive axillary nodes detected (baze Haberman's Survival Data Set) sa tačkama reza dobijenim na osnovu algoritma baziranog na entropiji

10. DODACI ZA ANALIZU BAZE CHALLENGER USA SPACE SHUTTLE O-RING

Bazu Challenger USA Space Shuttle O-Ring Data Set je donirao David Draper, University of California, Los Angeles. Eksplozija spejs šatla Challenger 28. januara 1986. godine je pokrenula naučnike i inženjere da ponovo preispitaju pouzdanost pogonskog sistema šatla. Eksplozija je bila povezana sa problemom sfernog zgloba koji je bio držač dva dela rakete. Svaki od šest zglobova je imao dva O-prstena (primarni i sekundarni) koji su imali mogućnost erozije tako da bi erozijom doveli do eksplozije. Tadašnji zaključak naučnika je bio da je samo temperatura bitan faktor erozije (manja temperatura, veći rizik), dok su zaključili da je pritisak nebitan. Zadatak je predvideti broj O-prstena koji će imati problem u odnosu na temperaturu, kada je temperatura ispod smrzavanja (ispod nule C stepena) [Challenger, 1993].

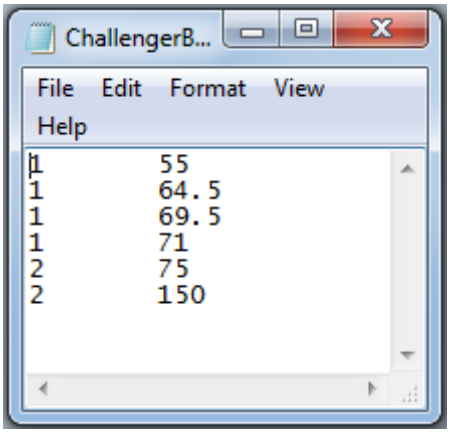
Informacije o atributima (preuzeto iz [Challenger, 1993]):

1. Number of O-rings at risk on a given flight
2. Number experiencing thermal distress
3. Launch temperature (degrees F)
4. Leak-check pressure (psi)
5. Temporal order of flight

Broj instanci je 23. U sistemu Rosetta urađena je diskretizacija ove baze samo za bitne atribute i dobijeni su sledeći rezultati:

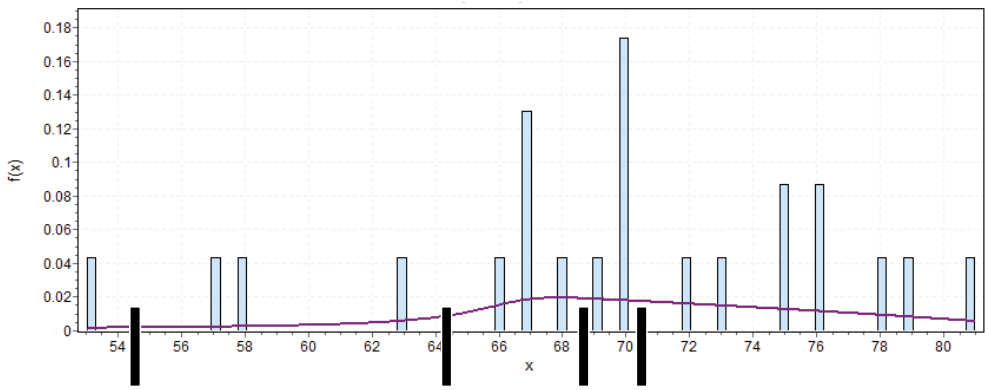
Algoritam maksimalne razberivosti

Na osnovu diskretizacije baze Challenger USA Space Shuttle O-Ring Data Set algoritmom maksimalne razberivosti, dobijene su tačke reza koje su prikazane na slici 227.

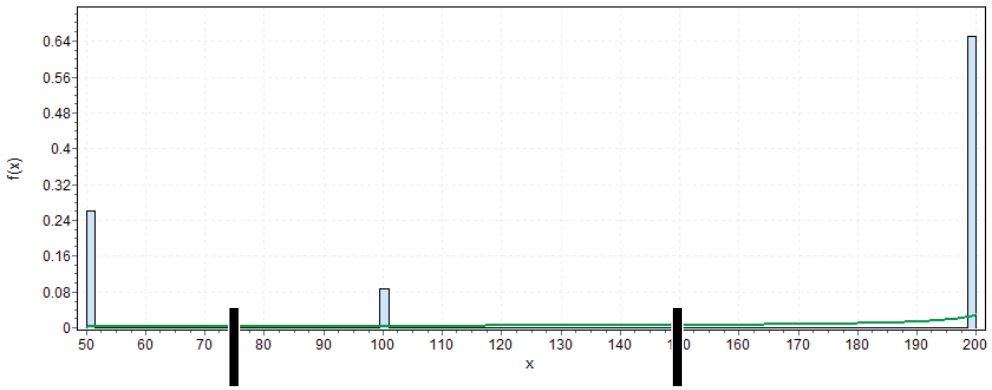


Slika 227. Tačke reza baze Challenger USA Space Shuttle O-Ring Data Set dobijene algoritmom maksimalne razberivosti

U odnosu na raspodelu podataka baze Challenger USA Space Shuttle O-Ring Data Set koja je analizirana softverom EasyFit, ove tačke reza (dobijene algoritmom maksimalne razberivosti) su prikazane na ordinatama širokim crnim vertikanim linijama (slike 228 - 229).



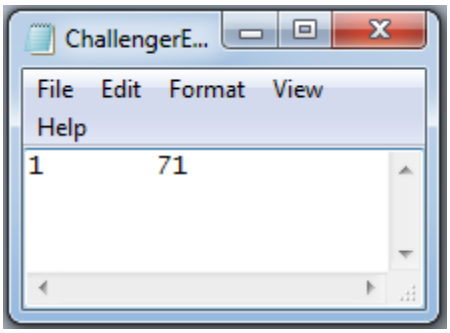
Slika 228. Raspodela podataka 3. atributa Launch temperature (degrees F, baze Challenger USA Space Shuttle O-Ring Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti



Slika 229. Raspodela podataka 4. atributa Leak-check pressure (psi, baze Challenger USA Space Shuttle O-Ring Data Set) sa tačkama reza dobijenim na osnovu algoritma maksimalne razberivosti

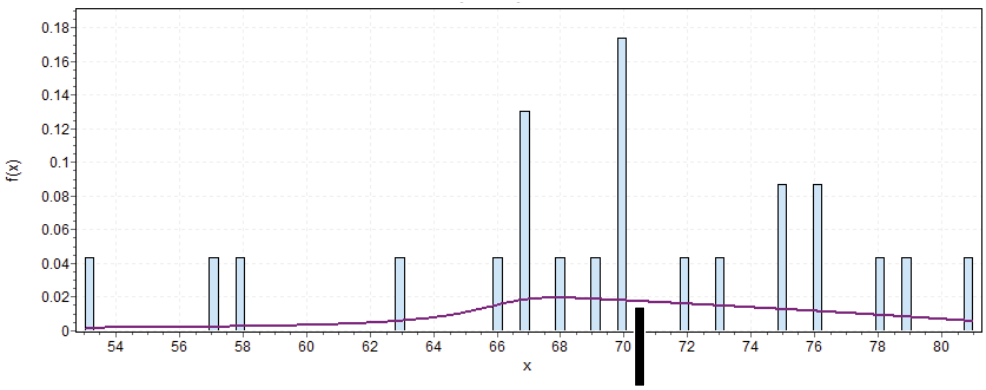
Algoritam baziran na entropiji

Na osnovu diskretizacije baze Challenger USA Space Shuttle O-Ring Data Set algoritmom baziranim na entropiji, dobijena je samo jedna tačka reza, koja je prikazana na slici 230.

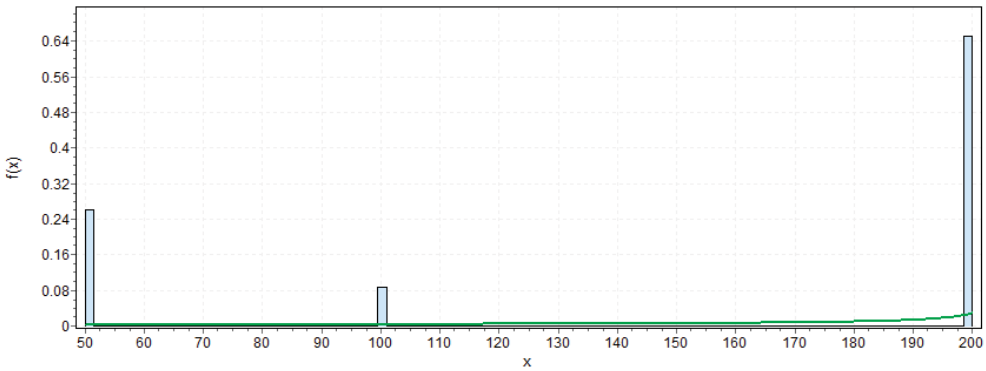


Slika 230. Samo jedna tačka reza baze Challenger USA Space Shuttle O-Ring Data Set dobijene algoritmom baziranim na entropiji

U odnosu na raspodelu podataka baze Challenger USA Space Shuttle O-Ring Data Set koja je analizirana softverom EasyFit, ova tačka reza prikazana je na ordinati širokom crnom vertikanom linijom (slike 231 - 232).



Slika 231. Raspodela podataka 3. atributa Launch temperature (degrees F, baze Challenger USA Space Shuttle O-Ring Data Set)) sa jednom tačkom reza dobijenom na osnovu algoritma baziranog na entropiji



Slika 232. Raspodela podataka 4. atributa Leak-check pressure (psi, baze Challenger USA Space Shuttle O-Ring Data Set) bez tačke reza na osnovu algoritma baziranog na entropiji