Diskretizacija podataka redukcijom tačaka reza

Višnja Ognjenović, Vladimir Brtka, Eleonora Brtka, Ivana Berković
Univerzitet u Novom Sadu, Tehnički fakultet "Mihajlo Pupin"
Zrenjanin
Srbija
visnjao@tfzr.uns.ac.rs, vbrtka@tfzr.uns.ac.rs, eleonorabrtka@gmail.com, berkovic@tfzr.uns.ac.rs

Sažetak—**U oblasti Data Mining-a mnogi obučavajući metodi mogu da rade samo sa diskretnim vrednostima atributima. Kontunialne vrednosti atributa mogu da se diskretizuju pomoću različitih metoda za diskretizaciju. Zamenom mnogih vrednosti atributa malim brojem intervala, originalni podaci se redukuju i pojednostavljuju. Rad predstavlja uticaj redukcije tačaka reza na rezultat klasifikacije u oblasti teorije grubih skupova.**

Ključne riječi - diskretizacija; tačke reza; klasifikacija; teorija grubih skupova

I. UVOD

U oblasti Data Mining-a mnogi obučavajući metodi (machine learning) mogu da rade samo sa diskretnim vrednostima atributa. Zbog toga pre machine learning procesa, neophodno je transformisati kontinualne vrednosti atributa u diskretne, konstituisane od skupa intervala. Ovaj process poznat kao diskretizacija podataka je esencijalni zadatak u preprocesiranju podataka, ne samo zbog toga što neki obučavajući metodi ne rade sa kontinualnim 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 algoritama, heuristike, a takođe i aproksimativne vrednosti rezultata diskretizacije.

Empirijski rezultati pokazuju da kvalitet klasifikacijskih metoda zavisi od algoritma diskretizacije koji se koristi [1]. Pošto 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 (cut) nad domenima atributa [2].

Postoji nekoliko podela na osnovu kojih je moguće klasifikovati algoritme za diskretizaciju. Prema [3],.[4] neke od osnovnih podela su sledeće:

- lokalna – globalna diskretizacija (local – global)
- dinamička – statička diskretizacija (dynamic – static)
- nadzirana – nenadzirana diskretizacija (supervised – unsupervised)
- univarijantna – multivarijantna diskretizacija (univariate – multivariate)
- deleća – objedinjujuća diskretizacija (splitting – merging, Top-Down and Bottom-up)
- direktna – inkrementalna diskretizacija (direct – incremental)
- po meri ocene diskretizacije (informacija, statistika, grubi skupovi, wrapper, binning)

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 to u nekim situacijama može da bude kontraproduktivno. Bitno je da se razume ceo proces diskretizacije kao prvi korak klasifikacije, ali i kompletan algoritam klasifikacije i dobijeni rezultati.

Na osnovu podela mogu se izvesti relacije između pojedinih diskretizacija, kao na primer da su sve dinamičke diskretizacije lokalne, ili da je nadgledana diskretizacija u odnosu na određeni algoritam lokalna.

U okviru teorije grubih skupova algoritam za diskretizaciju maksimalne razberivosti je dinamički i nadziran [2].

U ovom radu će se pokazati diskretizacija podataka u teoriji grubih skupova, tako da će razmatrati redukcija dobijenih tačaka reza. Time bi se smanjio broj intervala a to bi uticalo i na rezultat diskretizacije. Za diskretizaciju koristiće se algoritam maksimalne nerazberivosti (MD-heuristic algorithm) koji je u stvari greedy algoritam za određivanje minimalnog skupa pokrivanja (minimal set covering) objekata iz različitih klasa atributa odluke. Ovaj algoritam je implementiran u sistemu Rosetta koji će se koristiti za dobijanje početnih tačaka reza [5]. Redukcija dobijenih tačaka reza će se raditi na osnovu analize histograma podataka pojednih atributa, a za analizu histograma će se koristiti softver EasyFit [6]. Time bi se pokazalo kako se redukcijom tačaka reza na osnovu histograma, može poboljšati ukupan rezultat klasifikacije. Za klasifikaciju diskretizovanih podataka koristiće se rezultati

Rad je potpomognut sredstvima projekta CR 32044 „Razvoj softverskih alata za analizu i poboljšanje poslovnih procesa" koji finansira Ministarstvo za prosvetu, nauku i tehnološki razvoj Repulike Srbije

Džonsonovog algoritma za izračunavanje minimalnih prostih implikanti Bulove funkcije. Ovaj algoritam je takođe implementiran u sistemu Rosetta [5].

II. DISKRETIZACIJA U TEORIJI GRUBIH SKUPOVA

Teoriju grubih skupova je razvio Pawlak 1982 za analizu podataka. Osnovna namena grubih skupova je aproksimacija nepoznatih znanja preko poznatog znanja [7]. Za teoriju grubih skupova je bitno postojanje univerzuma koji sadrži objekte definisane pomoću vrednosti svojih atributa. Bazirana na principu nerazberivosti objekata i konceptu aproksimacije, ova teorija omogućuje prepoznavanje zavisnosti između atributa odluke i uslovnih atributa [8]. U ovom radu analiziraće se naknadna redukcija tačaka reza i njen uticaj na klsifikaciju u okviru teorije grubih skupova.

A. Osnove teorije grubih skupova

Podaci koji se analiziraju su tabelarno organizovani. U teoriji grubih skupova definisana je informaciona tabela [9]. 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:

$$Des_q(x) = [f(x,q_1), f(x,q_2), ..., f(x,q_m)] \tag{1}$$

koji definiše vrednosti atributa objekta x. Neka je sa P označen neprazan podskup skupa atributa Q. Definisana je relacija I_P nad U:

$$I_P = \{(x,y) \in U \times U : f(x,q) = f(y,q), \forall q \in P\} \tag{2}$$

Relacija (2) se zove relacija nerazberivosti, ili relacija nerazlikovanja (*indiscernibility relation*). 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 $I_P(x)$. Ako je $P = Q$, P-elementarni skupovi se nazivaju atomi (*atoms*).

Neka je S informaciona tabela, X neprazan podskup od U, a $\varnothing \neq P \subseteq Q$:

$$\underline{P}(X) = \{x \in U : I_P(x) \subseteq X\} \tag{3}$$

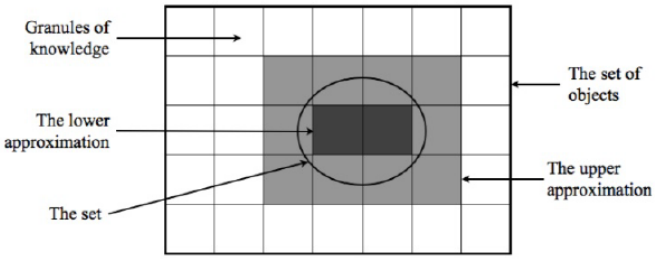
$$\overline{P}(X) = \bigcup_{x \in X} I_P(x) \tag{4}$$

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.

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

$$Bn_P(X) = \overline{P}(X) - \underline{P}(X) \tag{5}$$

Grafička interpretacija P-granice je prikazana na Sl. 1.



Slika 1. Osnovna ideja teorije grubih skupova, slika preuzeta iz [10]

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 $S = \langle U, (C \cup D), V, f \rangle$.

Generalizovana funkciju odluke ∂A (x_i) objekta x_i za skup $A \subseteq C$, definisana je kao skup klasa odluke po svim objektima u okviru klase ekvivalencije x_i [11].

$$\partial_A(x_i) = \{f(x_i, d) \mid x_i \in [x_i]_A\} \tag{6}$$

Za tabelu odluke se kaže da je konzistentna (consistent) ako je kardinalnost od ∂_A (x_i) jednaka 1 za sve objekte u univerzumu. Inače ako kardinalnost generalizovane funkcije odluke nije jednaka 1, tabela je nekonzistenta (inconsistent).

B. Osnovne definicije diskretizacije u teoriji grubih skupova

Diskretizacija kontinualnih podataka u okviru teorije grubih skupova je bazirana na definisanju skupa tačaka reza (set od 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 $l_c \le p_i < r_c$. Broj p_i pravi particiju objekata univerzuma U na dva disjunktna skupa U_l i U_r gde je

$$U_l = \{x_j \in U \mid f(x_j, c) \le p_i\} \tag{7}$$

i

$$U_r = \{x_j \in U \mid f(x_j, c) > p_i\} \tag{8}$$

Oba skupa U_1 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 $l_c \leqslant p_1 < p_2 < ... < p_k < r_c$.

Prema [11] 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

$$P = \bigcup_{c \in C} P_c \tag{9}$$

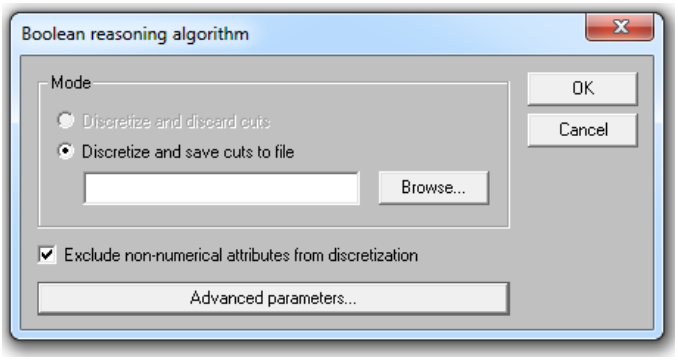
a funkcija f^P je definisana na sledeći način:

$$f^P(x_j, c) = \begin{cases} 0, & if \quad f(x_j, c) < p_1 \\ i, & if \ f(x_j, c) \in [p_i, p_{i+1}), 1 \le i \le k-1 \\ k, & if \quad f(x_j, c) \ge p_k \end{cases} \tag{10}$$

C. *Algoritam maksimalne razberivosti*

Algoritam maksimalne razberisvosti (MD-heuristic algorithm) koristi Boolean reasoning pristup [2] koji garantuje razberivost između objekata.

Ovaj algoritam je implementiran u sistemu Rosetta i na Sl.2 je prikazan njegov korisnički inferfejs:



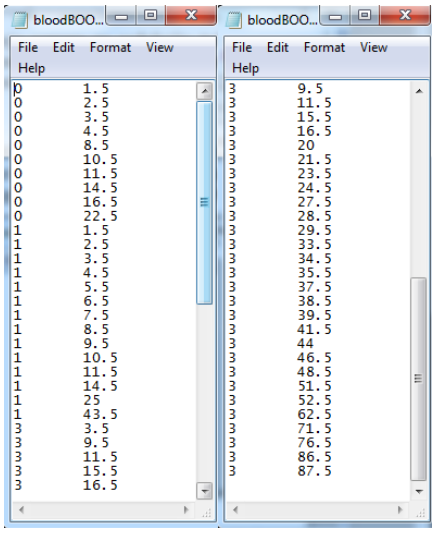
Slika 2. Korisnički interfejs MD algoritma

III. REDUKCIJA TAČAKA REZA

Glavna ideja vezana za redukciju tačaka reza dobijenih MD algoritmom je vezana za analizu odnosa tačaka reza i histograma vrednosti atributa.

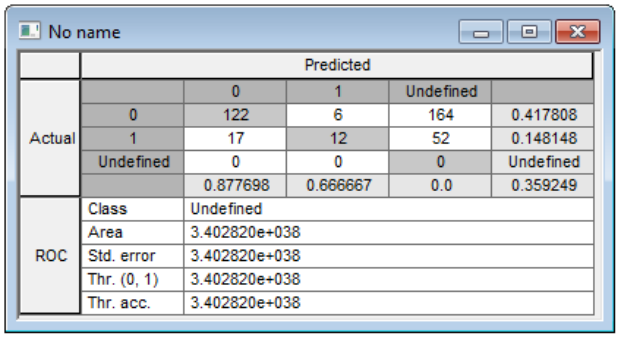
A. *Primer 1*

Za bazu Blood Transfusion Service Center Data Set [12], koja ima četiri uslovna atributa i jedan atribut odluke, na osnovu MD algoritma dobijene su sledeće tačke reza (Sl.3). Sa nulom je označen prvi atribut, sa jednicom drugi a sa brojem tri četvrti atribut. Treći atribut nije diskretizovan, odnosno MD-algoritam ga je izbacio pošto ne utiče na razberivost.



Slika 3. Tačke reza - rezultat MD algoritma

Rezultat klasifikacije na ovako diskretizovanim podacima je loš. Na Sl.4 je prikazana matrica konfuzije. Ukupna ocena je 35,92%.

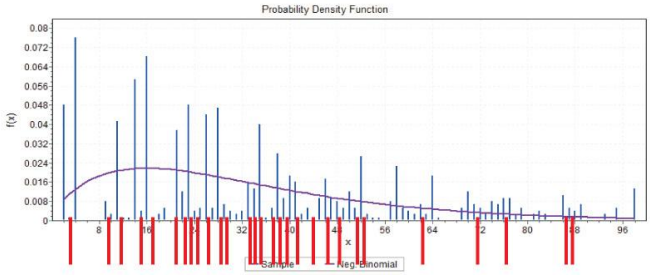


		Predicted			
		0	1	Undefined	
Actual	0	122	6	164	0.417808
	1	17	12	52	0.148148
	Undefined	0	0	0	Undefined
		0.877698	0.666667	0.0	0.359249

ROC	Class	Undefined
	Area	3.402820e+038
	Std. error	3.402820e+038
	Thr. (0, 1)	3.402820e+038
	Thr. acc.	3.402820e+038

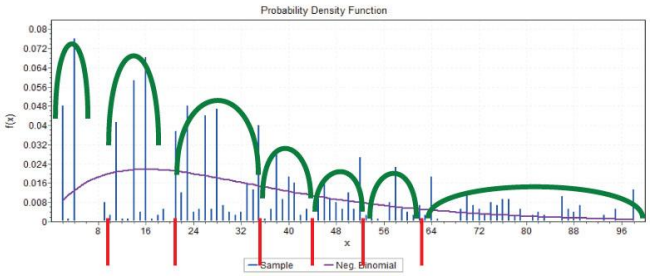
Slika 4. Rezultat klasifikacije pri diskretizaciji MD algritmom

Ako se pogleda histogram na primer četvrtog atributa, i njegove tačke reza, može se primetiti da one prate lokalne maksimume ili normalnu raspodelu oko lokalnog maksimuma (Sl. 5)



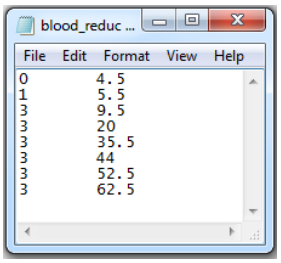
Slika 5. Tačke reza na histogramu

Ako se uradi redukcija tačaka reza tako da se uoče grupacije lokalnih maksimuma koje na jednom delu ukupnog intervala vrednosti atributa predstavljaju normalnu raspodelu na histogramu, onda grafički to može da izgleda kao na Sl. 6.

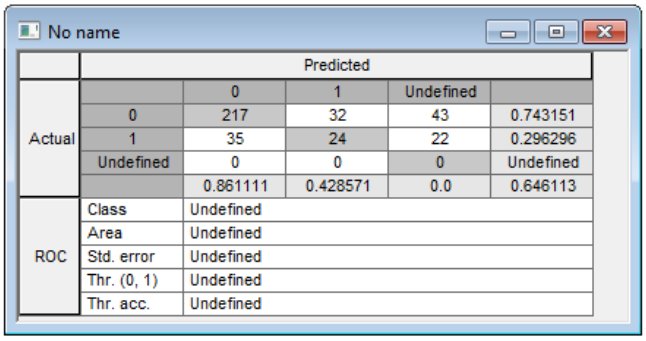


Slika 6. Tačke reza četvrtog atributa dobijene redukcijom na osnovu grupacija lokalnih maksimuma i raspodele na histogramu

Ako se istim postupkom redukuju tačke reza i kod ostalih atributa, onda se za tako redukovane tačke reza koje su ručno unete u sistem Rosetta (Sl. 7), dobija rezultat klasifikacije prikazan na Sl. 8.



Slika 7. Redukovan skup tačaka reza na osnovu grupacija lokalnih maksimuma



Slika 8. Rezultat klasifikacije za redukovan skup tačaka reza na osnovu grupacija lokalnih maksimuma

Na osnovu rezultata matrice konfuzije može se videti da se na ovaj način diskretizovana tabela može bolje klasifikovati 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:

- dobijeno je 266 pravila (na osnovu podataka diskretizovanih MD algoritmom – algoritmom maksimalne razberivosti) od kojih njih 11 ima veznik OR u THEN delu pravila (Sl. 9).
- dobijeno je 118 pravila (na osnovu redukcije tačaka reza) od kojih njih 30 ima veznik OR u THEN delu pravila (Sl. 10).

26	2([2, 3)) AND 50([7, 8)) AND 98([28, 29)) => 1(1)
27	2([2, 3)) AND 50([8, 9)) AND 98([35, 36)) => 1(1)
28	2([4, 5)) AND 50([5, 6)) AND 98([16, 17)) => 1(1)
29	2([2, 3)) AND 50([3, 4)) AND 98([4, 10)) => 1(1)
30	2([3, 4)) AND 50([15, 25)) AND 98([72, 77)) => 1(0)
31	2([2, 3)) AND 50([4, 5)) AND 98([12, 16)) => 1(0) OR 1(1)
32	2([*, 2)) AND 50([9, 10)) AND 98([49, 52)) => 1(0)
33	2([2, 3)) AND 50([4, 5)) AND 98([16, 17)) => 1(0) OR 1(1)
34	2([2, 3)) AND 50([2, 3)) AND 98([4, 10)) => 1(0) OR 1(1)
35	2([2, 3)) AND 50([6, 7)) AND 98([28, 29)) => 1(1)

Slika 9. Deo pravila dobijenih nad podacima koji su diskretizovani MD algoritmom

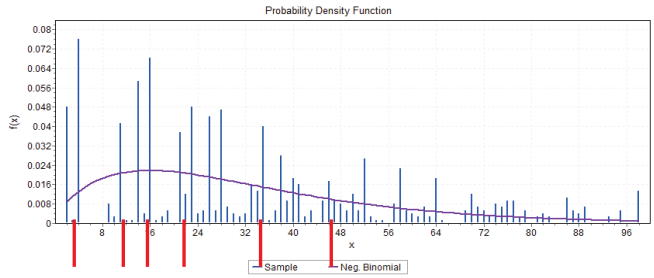
26	2([*, 5)) AND 12500(2250) AND 98([36, 44)) => 1(1)
27	2([5, *)) AND 12500(1750) AND 98([20, 36)) => 1(0) OR 1(1)
28	2([*, 5)) AND 12500(1250) AND 98([20, 36)) => 1(0) OR 1(1)
29	2([5, *)) AND 12500(4250) AND 98([63, *)) => 1(0) OR 1(1)
30	2([*, 5)) AND 12500(750) AND 98([10, 20)) => 1(0) OR 1(1)
31	2([*, 5)) AND 12500(2500) AND 98([63, *)) => 1(0)
32	2([*, 5)) AND 12500(2000) AND 98([36, 44)) => 1(1)
33	2([*, 5)) AND 12500(3000) AND 98([63, *)) => 1(0) OR 1(1)
34	2([5, *)) AND 12500(6000) AND 98([63, *)) => 1(0)
35	2([*, 5)) AND 12500(2750) AND 98([53, 63)) => 1(0)

Slika 10. Deo pravila dobijenih nad podacima koji su diskretizovani redukovanim skupom tačaka reza sa Sl. 7

Povećanjem broja pravila koja imaju OR smanjuje se razberivost, odnosno isti podaci se klasifikuju na dva različita načina. Zbog toga matrica konfuzije može da ima dobar rezultat a da se u stvari ne zna tačna odluka za konkretan objekat.

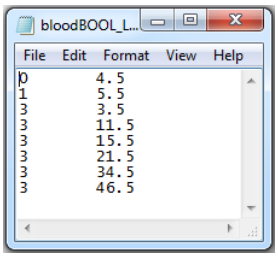
B. Primer 2

Da bi se pokazala značajnost grupisanja lokalnih maksimuma sa Sl. 6, u odnosu na neku drugu redukciju tačaka, u ovom primeru će se ponovo poći od iste baze i od tačaka reza dobijenih Bulovim algoritmom maksimalne razberivosti (Sl. 3 i Sl. 5). U odnosu na redukovan skup tačaka reza sa Sl. 7, izmeniće se samo tačke reza četvrtog atributa. Namerno će se zaobići grupisanje oko lokalnih maksimuma kod četvrtog atributa, odnosno uzeće se tačke koje se nalaze unutar grupacija lokalnih maksimuma. Takav izbor tačaka reza četvrtog atributa je prikazan na histogramu na Sl. 11.

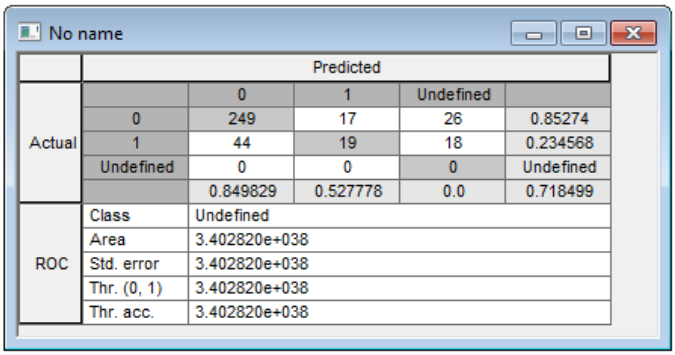


Slika 11. Izmenjen skup tačaka reza kod četvrtog atributa tako da nije poštovano grupisanje oko lokalnih maksimuma

Za tako izabrane tačke reza četvrtog atributa, koje su ručno unete u sistem Rosetta (Sl. 12), dobija se rezultat klasifikacije prikazan na Sl. 13.



Slika 12. Izmenjen skup tačaka reza - za četvrti atribut tačke reza sa Sl. 11



		Predicted			
		0	1	Undefined	
Actual	0	249	17	26	0.85274
	1	44	19	18	0.234568
	Undefined	0	0	0	Undefined
		0.849829	0.527778	0.0	0.718499

ROC	Class	Undefined
	Area	3.402820e+038
	Std. error	3.402820e+038
	Thr. (0, 1)	3.402820e+038
	Thr. acc.	3.402820e+038

Slika 13. Rezultat klasifikacije za redukovan skup tačaka reza na osnovu negrupisanja lokalnih maksimuma

C. Komparacija

Ono što deluje kao napredak u Primeru 2 je u stvari problem jer je klasifikacija dobijena značajno većim povećanjem broja pravila koja u THEN delu imaju OR. Na bazi diskretizacije iz Primera 2, dobijeno je ukupno 117 pravila od kojih njih 38 ima operator OR u THEN delu.

U odnosu na redukciju tačaka reza iz Primera 1, prva dva atributa imaju iste tačke reza, treći atribut nije diskretizovan, dok je za četvrti atribut uzet isti broj tačaka reza ali je u Primeru 1 poštovan princip grupisanja lokalnih maksimuma, a u Primeru 2 nije. U Primeru 1 dobijeno je 118 pravila, a u Primeru 2, 117 pravila. Ono što je značajna razlika je što je u Primeru 1, operator OR u THEN delu pravila imalo 30 pravila, dok je u Primeru 2, čak 38 pravila imalo operator OR u THEN delu pravila. Za izmenu tačaka reza kod samo jednog atributa i za isti broja tačaka reza kao i ujednačene veličine intervala dobijenih tačkama reza, 8 pravila više, čini značajnu razliku.

Time se u Primeru 2 u većoj meri nego u Primeru 1 (sa tačkama reza dobijenim redukcijom) narušava razberivost.

D. Metod redukcije tačaka reza

Radom u sistemu Rosetta primećeno je da se u zavisnosti od vrste podataka dobijaju određeni rezultati. Istraživanjem histograma podataka, kao i raspodela nad histogramima, potvrđeno je da redukcija tačaka reza dobijenih MD algoritmom zavisi od sledećeg:

- u kojoj meri podaci predstavljeni histogramom odgovaraju ili ne odgovaraju normalnoj raspodeli ili normalnim raspodelama na pojedinim delovima histograma – na osnovu toga redukcija tačaka reza može da se uradi kao u Primeru 1;
- koliki je broj tačaka reza – kod malog broja tačaka reza, redukcijom se uglavnom dobijaju lošiji rezultati a što je broj tačaka reza veći, rezultati redukcije su bolji;
- koliki je rezultat klasifikacije – kod dobrog rezultata klasifikacije izbacivanjem tačke reza koja se na histogramu već nalazi u podintervalu okoline lokalnog maksimuma, u velikom broju slučajeva dobija se isti rezultat.

Ovo su samo osnovni parametri koji mogu da pomognu u situacijama kada je rezultat klasifikacije loš. U slučaju kada je rezultat klasifikacije dobar, redukcijom tačaka reza može da se naruši razberivost a time dobije lošiji rezultat klasifikacije.

IV. ZAKLJUČAK

U radu je pokazano na koji način histogram vrednosti atributa može da utiče na izbor tačaka reza za redukciju. Na osnovu odgovarajućih primera pokazan je problem generisanja velikog broja tačaka reza nad podacima koji imaju veći broj lokalnih maksimuma na histogramu. Redukcijom onih tačka koje se nalaze u okolini lokalnih maksimuma, pored smanjenja broja tačaka reza, dobija se bolji rezultat klasifikacije uz manje smanjenje razberivosti.

Ovakav metod može da pomogne ekspertu da bolje razume uticaj diskretizacije na klasifikaciju podataka, kao i da izbegne lošu redukciju tačaka reza.

LITERATURA

[1] J. Gama, L. Torgo, C. Soares, "Dynamic Discretization of Continuons Attributes", www.liaad.up.pt/~ltorgo/Papers/DDCA.ps.gz

[2] HS Nguyen, Approximate boolean reasoning: foundations and applications in data mining, Transactions on rough sets V, 334-506, 2006.

[3] Sergio Ramırez-Gallego, Salvador Garcıa, Hector Mourino-Talın, David Martınez-Rego, Veronica Bolon-Canedo, Amparo Alonso-Betanzos, Jose Manuel Benıtez, Francisco Herrera, Data Discretization: Taxonomy and Big Data Challenge, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 6, Issue 1, pages 5–21, January/February 2016.

[4] Stephen D. Bay, 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.

[5] Øhrn, A.: Rosetta Technical Reference Manual (1999), http://www.idi.ntnu.no/_aleks/rosetta

[6] EasyFit - Distribution Fitting Software, http://www.mathwave.com/easyfit-distribution-fitting.html

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

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

[9] Komorowski J., Pawlak Z., Polkowski L., Skowron A., "Rough Sets: A Tutorial", http://citeseer.ist.psu.edu/komorowski98rough.html, 1998.

[10] Gloria Virginia, Lexicon-based Document Representation, Fundamenta Informaticae 124 (2013) 27–46

[11] Srilatha Chebrolu, Sriram G. Sanjeevi, 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

[12] Blake, C.L., Merz, C.J.: UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/

ABSTRACT

In the Data Mining field, many learning methods can handle only discrete attributes. Continuous features in the data can be discretized using different discretization methods. Replacing numerous values of a continuous attribute by a small number of intervals thereby reduces and simplifies the original data. Paper presents the impact of reducing the cuts on the result of classification in the rough set theory.

DATA DISCRETIZATION BY REDUCTION OF CUTS

Visnja Ognjenovic, Vladimir Brtka, Eleonora Brtka, Ivana Berkovic