Vladislav Mišković,
potpukovnik, dipl. inž.
Centar za informatičku podršku pozadine GŠ VJ,
Beograd

PRIMENA SISTEMA INDUKTIVNOG MAŠINSKOG UČENJA U INTELIGENTNOJ ANALIZI PODATAKA

UDC: 007.52 : 681.3.06

Rezime:

U ovom radu razmatrana je primena metoda induktivnog mašinskog učenja u inteligentnoj analizi podataka (data mining, knowledge discovery in databases). Prikazane su i osnovne karakteristike sopstvenog nekomercijalnog sistema za inteligentnu analizu podataka, kao i eksperimentalno poređenje njegove komponente za induktivno učenje pravila sa najpoznatijim sistemima iz te kategorije.

Ključne reči: istraživanje podataka, induktivno mašinsko učenje, softver.

APPLICATION OF INDUCTIVE MACHINE LEARNING IN DATA MINING

Summary:

This paper considers the application of inductive machine learning methods in data mining and knowledge discovery in databases tasks. Some characteristics of the author's noncomercial System for Data Mining and Knowledge Discovery are presented as Well as the experimental comparison of its inductive rules learning component with some best known systems from the same category.

Key words: data mining, knowledge discovery, inductive machine learning, software, comparison.

Uvod

Prikupljanje velikih količina raznovrsnih podataka u savremenim informacionim sistemima stvorilo je potrebu za programima koji mogu efikasno da im pristupe i izdvoje one koji su korisni za određenu svrhu [1, 2].

Primeri procesa koji generišu veliki broj podataka su razne evidencije o ljudima (matične, pravosudne, krivične, finansijske, zdravstvene, obrazovne i slično), proizvodnji (nadgledanje procesa, upravljanje kvalitetom), prodaji (klijenti) i eksploataciji proizvoda (dijagnostika, servisiranje). Struktura, format i smisao podataka su vrlo raznovrsni, i najčešće se ne uklapaju u standardne matematičke modele, tako da je njihova analiza klasičnim statističkim metodama veoma složena ili čak nemoguća.

Sistemi za *inteligentnu analizu podataka* (data mining, knowledge discovery in databases) jesu alati koji mogu da analiziraju sadržaj velikih baza podataka i ustanove određene zakonitosti u njima. Dobijeno znanje se zatim koristi za donošenje odluka zasnovanih na znanju, npr.

u sistemima za dijagnostiku oboljenja ili tehničkih sistema.

Inteligentna analiza podataka

Inteligentna analiza podataka je autorov zajednički naziv za dva pojma – *istraživanje podataka* (data mining) i *otkrivanje znanja* u bazama podataka (knowledge discovery in databases – KDD), koji se ponekad u literaturi poistovećuju [2].

Prema [1, 3], istraživanje podataka obuhvata primenu metoda mašinskog učenja i drugih metoda, za pronalaženje svih uzoraka u posmatranim podacima („enumeration of patterns over the data"), dok se otkrivanje znanja odnosi na celokupan životni ciklus analize podataka, od identifikacije ciljeva analize, prikupljanja i organizacije sirovih podataka do generisanja potencijalno korisnog znanja, njegove interpretacije i testiranja.

Prvi pristup [1, 3] istraživanje podataka definiše kao „izdvajanje uzoraka ili modela iz posmatranih podataka" i razmatra ga kao deo procesa otkrivanja znanja. Pod otkrivanjem znanja podrazumeva se „netrivijalan proces identifikacije novih, potencijalno korisnih i obavezno razumljivih uzoraka u podacima", koji obuhvata:

– razvoj i razumevanje primene i ciljeva procesa istraživanja podataka;
– prikupljanje i odabir podataka;
– objedinjavanje i proveru izabranog skupa podataka;
– čišćenje, pretprocesiranje i transformaciju;
– razvoj modela i postavljanje hipoteza;
– izbor pogodnog algoritma istraživanja podataka;
– istraživanje podataka (data mining);
– interpretaciju rezultata i vizualizaciju;
– testiranje i verifikaciju rezultata;
– upotrebu i održavanje otkrivenog znanja.

Prema [3], istraživanje podataka se procenjuje na 15–25% ukupnih napora u celom procesu otkrivanja znanja. Prema drugom pristupu [2] koristi se jedinstveni termin „istraživanje podataka" za celokupan proces i definiše kao „korišćenje istorijskih podataka za otkrivanje pravilnosti i poboljšanje budućih odluka".

U analizi podataka najčešće se koristi model problema zasnovan na *objektima* i *atributima* koji ih opisuju. Svaka konkretna reprezentacija objekta određena je kombinacija *vrednosti* atributa i obično se, pojednostavljeno, takođe naziva objektom. Slični objekti (po nekom kriterijumu) mogu se grupisati u *klase,* dok različiti objekti pripadaju različitim klasama. Formiranje klasa redukuje kompleksnost podataka i otkriva strukturu u podacima. Savremeni alati za analizu automatizuju ovaj postupak.

U analizi podataka koriste se tradicionalne metode (regresivna analiza, klaster-analiza, numerička taksonomija, stohastičko modeliranje, itd.), kao i razne *metode mašinskog učenja* (fuzzy logika, neuronske mreže, sistemi za učenje produkcionih pravila i stabala odlučivanja, itd.). Suštinska razlika je u tome što su tradicionalne metode okrenute kvantitativnim osobinama podataka kao celine, a rezultat analize kreira sam analitičar.

Rezultat metoda mašinskog učenja jesu logičke zakonitosti i kvalitativni opisi, koje kreira sam alat. Takvi opisi mogu da sadrže više oblika izražavanja istovremeno, na primer, logički, matematički,

statistički i grafički. Osnovni uslov je da budu lako razumljivi analitičarima i ekspertima koji se bave razmatranim područjem, tj. da zadovoljavaju „princip razumljivosti" (principle of comprehensibility) [1].

Induktivno mašinsko učenje na osnovu primera

Mogućnost učenja je jedan od osnovnih pokazatelja inteligentnog ponašanja. Izučavanje i računarsko modeliranje procesa učenja predstavlja predmet proučavanja na području mašinskog učenja.

Prema [4] *mašinsko učenje* (machine learning) jeste skup procesa, gde spadaju:

– prikupljanje novog deklarativnog znanja;

– razvoj i usavršavanje motornih i saznajnih sposobnosti kroz praksu;

– strukturisanje postojećeg znanja;

– otkrivanje novih činjenica i teorija posmatranjem i aktivnim eksperimentisanjem.

Učenje se može posmatrati kroz dve osnovne forme:

– prikupljanje znanja (knowledge acquisition), tj. učenje nove (simboličke) informacije, tako da se ona može primeniti u efektivnom smislu (primer: tako čovek uči fiziku);

– uvežbavanje (training), tj. poboljšavanje već stečenog znanja, bilo mentalne, bilo motorne koordinacije, putem praktičnog ponavljanja i korekcije odstupanja od željenog ponašanja (primer: tako čovek uči veštine – vožnju bicikla ili sviranje na klaviru, kada prikupljanje znanja predstavlja tek prvu fazu učenja).

Smatra se da učenje kod čoveka predstavlja mešavinu obe forme, s tim da mentalne aktivnosti potenciraju prvu formu, a motorne aktivnosti u većoj meri drugu formu učenja.

Sistemi mašinskog učenja najčešće se dele prema odabranoj strategiji učenja, načinu predstavljanja znanja i području primene [4]. Podela prema strategiji učenja odnosi se na potrebnu količinu zaključivanja sistema (u ulozi učenika) nad podacima koje dobija iz okruženja (u ulozi učitelja):

– učenje memorisanjem (rote learning), kada nema zaključivanja ni transformacije znanja (obično programiranje i upotreba primitivnih baza podataka);

– učenje na osnovu rečenog (learning by being told), kada se deklarativno znanje transformiše u internu formu predstavljanja i integriše sa postojećim znanjem. Takvo znanje sistem upotrebljava bez dodatnog programiranja;

– učenje po analogiji (learning by analogy), kada se zahteva veća količina zaključivanja, jer se novo znanje stiče transformacijom i proširivanjem postojećeg znanja u takvom obliku da se može upotrebiti za rešavanje novih problema, koji su u određenoj meri sličnі već rešenim;

– učenje na osnovu primera (learning by examples), koje zahteva induktivno zaključivanje. Analizom i generalizacijom rešenih primera i kontraprimera neke klase pojava (pojma) dolazi se do pravila, teorije ili opisa pojma, koji obuhvata sve primere i nijedan kontraprimer. Ovakve metode se najviše istražuju, a dalje se mogu klasifikovati prema izboru primera, izvoru primera i načinu upotrebe primera;

– učenje posmatranjem i samostalnim otkrivanjem (learning by observation and discovery), ili učenje bez učitelja, zahteva najveću količinu zaključivanja, jer sistem mora samostalno da otkriva

nove i značajne klase objekata (pojmove), postavlja hipoteze i proverava ih, te stvara teorije.

Poslednje dve strategije učenja (učenje na osnovu primera i učenje posmatranjem i samostalnim otkrivanjem) poznate su pod imenom *induktivno mašinsko učenje*.

Druga podela sistema mašinskog učenja zasniva se na iskustvu koje navodi na zaključak da postoje znanja koja se ne mogu eksplicitno izraziti [5], pa se razlikuju sistemi sa eksplicitnim znanjem, predstavljenim logikom, pravilima, frejmovima i sličnim načinima i sistemi sa implicitnim (distribuiranim) znanjem, npr. u neuronskim mrežama.

Prema području primene razlikuju se sistemi opšte namene i specijalizovani sistemi mašinskog učenja za posebna uska područja primene (npr. za razvoj ekspertskih sistema, poljoprivredu, hemiju, programiranje računara i robota, edukaciju, matematiku, prepoznavanje slike i govora, razumevanje prirodnog jezika, planiranje, predviđanje, igre, složene sisteme naoružanja, itd.).

Pregled poznatijih komercijalnih sistema za inteligentnu analizu podataka

Od velikog broja sistema za inteligentnu analizu podataka, radi ilustracije, prikazane su (tabela 1) osnovne karakteristike nekoliko značajnijih [3]:

Vidi se da svi sistemi koriste širok dijapazon različitih metoda induktivnog učenja kao osnovni alat za inteligentnu analizu podataka. Svi sistemi imaju ugrađenu neku od metoda za induktivno učenje simboličkih opisa (stabala odlučivanja ili produkcionih pravila), za šta koriste najpoznatije algoritme iz te kategorije (npr., ID3 [4], C4 [11], CN2 [13], CART).

Primer sopstvenog sistema za inteligentnu analizu podataka

Prvi prototip sistema za inteligentnu analizu podataka pod nazivom Empiric, autor je razvio na Katedri za računarsku tehniku VVTŠ KoV JNA (verzija za DOS). U ovom radu koristi se nova verzija za Windows okruženje.

Sistem za induktivno učenje Empiric

Sistem za induktivno učenje na osnovu primera zamišljen je kao alat u zadacima istraživanja podataka (data mining), odnosno otkrivanja znanja (knowledge discovery). Sastoji se od više podsistema koji obezbeđuju minimalni skup alata za inteligentnu analizu podataka metodama induktivnog mašinskog učenja:

– editor primera, namenjen za unos i ažuriranje modela problema i primera;

– podsistem za induktivno učenje bez učitelja: generator jednostavnih klasifikacija, realizovan algoritmom partitivnog grupisanja (partitional clustering) i generator hijerarhije klasa, realizovan algoritmom hijerarhijskog grupisanja (hierarhycal clustering);

– podsistem za induktivno učenje pravila na osnovu primera (induction of conjuctive rules);

– vizualizator, realizovan algoritmom za nelinearnu projekciju prostora primera (nonlinear mapping) u dve dimenzije, radi prikaza strukture skupa primera geometrijskim rasporedom tačaka na površini. Omogućava praćenje rada algoritama za generisanje klasifikacija.

Tabela 1

Pregled poznatijih sistema za inteligentnu analizu podataka

Red. br.	Naziv	Proizvođač	Operativni sistem	Format podataka	Obim podataka	Induktivne metode[1]
1.	Clementine	Integral Solutions Ltd.	Unix, WinNT	Tekst, Informix, Oracle, Sybase, Excel	$<10^6$	NN, DT, RI
2.	Darwin	Thinking Machines[2]	Unix[2]	Tekst, Oracle, Sybase	$>10^6$	NN, GA, DT
3.	Intelligent Miner	IBM	Unix, WinXX, OS/2	Tekst, Oracle, Sybase	$>10^6$	NN, DT
4.	MineSet	Silicon Graphics	Unix	Tekst, Informix, Oracle, Sybase	$>10^6$	St, DT, RI
5.	PolyAnalyst	Megaputer	WinXX, OS/2	Tekst, Informix, Oracle, Sybase, Access, Excel	$<10^6$	GA, St, RI

[1] NN – neuronske mreže, GA – genetički algoritmi, DT – stabla odlučivanja, RI – indukcija pravila, St – statističke metode
[2] Sada je postao deo sistema Oracle, pa je raspoloživ na svim platformama

Svi podsistemi prilagođeni su jedinstvenom modelu predstavljanja znanja, tzv. atributnom modelu sa više tipova atributa. Diskretni atributi ugrađeni su u tri varijante: nominalni (neuređen skup vrednosti), linearni (uređen skup vrednosti) i strukturni (parcijalno uređen skup vrednosti).

Osnovne karakteristike sistema Empiric

Osnovne karakteristike sistema u celini su:

– formira uniformni model primera sa više tipova atributa;

– dozvoljava nepoznate vrednosti u primerima za sve funkcije sistema;

– vizuelno prikazuje klasifikaciju radi boljeg uvida u strukturu i rad generatora klasifikacija;

– omogućava odabir primera za učenje i primera za testiranje pravila (slučajno i sekvencijalno);

– pri učenju pravila vrši automatsku dinamičku diskretizaciju kontinualnih atributa;

– generisano znanje predstavlja u obliku konjuktivnih pravila;

– omogućava dodatno povećanje tačnosti predviđanja učenjem višestrukih modela (bagging), što se u ovom radu ne razmatra.

U podsistem za induktivno učenje pravila sistema Empiric ugrađeno je više različitih mera za ocenu kvaliteta pravila u fazi njihovog formiranja:

– informativnost ili prirast informacije (information gain, ID3);

– relativna informativnost (gain ratio, C4);

– entropija (entropy, CN2);

– mera nečistoće klasifikacije (gini index, CART);

– mera logičke zasnovanosti (logical sufficiency content, HYDRA);

– mera prirasta informacije (Q-measure, AQ18).

U zagradi je naveden po jedan svetski poznat algoritam učenja koji tipično koristi odgovarajuću meru. U sistemu Empiric meru kvaliteta pravila bira korisnik pri aktiviranju algoritma učenja pravila, zajedno sa načinom izdvajanja i brojem primera za učenje i testiranje naučenih pravila.

Rezultati testiranja podsistema induktivnog učenja pravila na nezavisnim testovima

Izvršeno je testiranje podsistema za induktivno učenje pravila na skupu od 10 problema (tabela 2) iz baze podataka problema mašinskog učenja na University of California at Irvine [8]. Odabrani su problemi iz više različitih područja primene, sa različitim brojem primera, vrstom i brojem atributa, koji se često citiraju u referentnoj literaturi.

Tačnost predviđanja (predictive accuracy), koja se ponekad neprecizno naziva klasifikacijska tačnost, osnovni je pokazatelj performansi sistema induktivnog učenja. Predstavlja procenat uspešnosti klasifikacije novih, nerazmatranih primera korišćenjem naučenih pravila.

U tabeli 3 dato je poređenje tačnosti predviđanja podsistema za učenje pravila sistema Empiric sa najboljim rezultatima drugih sistema koji su pronađeni u literaturi, za svaki od navedenih problema [8, 9, 10, 11, 12]. Poređenje je izvršeno sa programima induktivnog učenja koji daju simbolički opis rezultata učenja, odnosno zadovoljavaju „princip razumljivosti".

Sistem Empiric je testiran pod istim uslovima (u pogledu načina izbora i broja primera u skupu za učenje/testiranje) za svaku od ugrađenih mera kvaliteta pravila i uzet je najbolji (Best) rezultat za poređenje.

Vidi se da je, po tačnosti predviđanja, podsistem za induktivno učenje pravila sistema Empiric uporediv sa algoritmima koji su ugrađeni u druge sisteme za inteligentnu analizu podataka, dakle, sa svetskim standardima u ovoj oblasti.

Dobijeni rezultati su relevantni samo za ponašanje algoritma koji se primenjuju na referentne i srodne probleme. Prema [7] ne mogu se automatski očekivati iste performanse na drugim problemima, zbog „zakona očuvanja" generalizacionih performansi, po kojem:

– ne postoji apsolutno najbolji algoritam učenja za sve probleme,

– algoritam učenja može biti bolji od drugog u jednoj situaciji samo na

Tabela 2

Pregled referentnih problema mašinskog učenja

Red. br.	Problem	Broj primera	Broj atributa		Broj klasa	% većinske klase	Ispušt. vredn.
			Diskretnih	Kontinualnih			
Problemi opisani samo diskretnim atributima							
1.	Kr-vs-kp (Chess)	3.196	36	–	2	52,22%	–
2.	Splice (DNA)	3.190	60	–	2	51,88%	–
Problemi opisani samo kontinualnim atributima							
3.	Iris Plant	150	–	4	3	33,33%	–
4.	Ionosphere	351	–	34	2	64,10%	–
5.	Pima Diabetes	768	–	8	2	65,10%	–
6.	Shuttle	58.000	–	9	7	78,60%	–
Problemi opisani sa obe vrste atributa i ispuštenim vrednostima							
7.	Hepatitis	155	13	6	2	79,35%	da
8.	Annealing	898	32	6	6	76,19%	da
9.	Thyroid Disease	3.772	22	7	3	95,68%	da
10.	Adult	48.842	8	6	2	76,07%	da

Tabela 3

Poređenje tačnosti predviđanja sa rezultatima iz literature za programe koji daju simboličke opise naučenih pravila

Red. br.	Problem	% več. klase	Način testiranja	Referenca/ algoritam		Najbolji rezultat %	Empiric (Best) %
Problemi opisani samo diskretnim atributima							
1.	Kr-vs-kp (Chess)	52,22%	10×(2.130:1.066 Random)	[12]	C4.5	99,10±0,00	96,90±0,63
2.	Splice (DNA)	51,88%	10×(2.000:1.190 Random)	[12]	ID3	94±0,00	87,01±1,34
Problemi opisani samo kontinualnim atributima							
3.	Iris Plant	33,33%	10×(70%:30% Random)	[9]	Assistant-R	95,40±2,60	94,44±3,62
4.	Ionosphere	64,10%	1×(200:151 Sequential)	[8]	C4	94,00±0,00	97,35±0,00
5.	Pima Diabetes	65,10%	50×(67%:33% Random)	[10]	CN2	73,60±2,40	72,89±3,57
6.	Shuttle 43.500	78,60%	1×(43.500:14.500 Sequential)	[8]	ID3	99,99±0,00	99,95±0,00
Problemi opisani sa obe vrste atributa i ispuštenim vrednostima							
7.	Hepatitis	79,35%	10×(70%:30% Random)	[9]	Assistant-R	83,00±3,50	81,74±3,79
8.	Annealing	76,19%	50×(67%:33% Random)	[10]	RISE	97,40±0,90	99,15±0,55
9.	Thyroid Disease	95,68%	10×(2.800:972 Random)	[11]	C4.5	99,52±0,10	96,22±0,41
10.	Adult 32.561	76,07%	1×(32.561:16.281 Sequential)	[8]	C4.5	85,50±0,00	80,74±0,00

račun gubitka performansi u nekim drugim situacijama,

– srednja klasifikacija tačnost algoritma u odnosu na sve matematički moguće probleme je konstantna i ne zavisi od algoritma.

Zbog toga se performanse algoritama učenja (pre svega tačnost predviđanja) mere i poboljšavaju u odnosu na određene stvarne probleme, a teži se tome da se gubitak performansi odrazi na probleme koji se nikad neće javiti u praksi.

Primer inteligentne analize podataka

Za ilustraciju procesa inteligentne analize metodama induktivnog mašinskog učenja upotrebiće se sistem Empiric i baza podataka o 155 modela automobila,[1] gde je svaki slog ili red tabele jedan primer konkretnog modela automobila. Traže se interesantne relacije u skupu objekata, koji su opisani svojstvima kao što su broj pređenih milja sa 1 galonom goriva (MPG), broj cilindara (cylinders) i godina proizvodnje (year), slika 1.

Kada se skup primera prikaže u prozoru vizuelizatora, uočava se izraženo grupisanje primera u nekoliko grupa (slika 2).

Interesantno je ispitati grupisanje na dve grupe (dijagonalno raspoređene) pomoću programa za generisanje particija. Rezultat rada algoritma grupisanja vidi se na slici 3, mada nije obavezno da se grupisanje koje generiše program poklopi sa grupisanjem na vizuelnom prikazu.

Koje su karakteristike ova dva podskupa podataka? Odgovor daje program induktivnog mašinskog učenja, koji ima zadatak da logički opiše ovo grupisanje primera. Rezultat je objašnjenje generisane klasifikacije na slici 4.

Ovaj skup pravila često se izražava u obliku pravila odlučivanja (decision rules) koja se često koriste u ekspertnim sistemima, slika 5.

[1] Ilustrativan primer iz statističkog paketa StatGraf 2.0

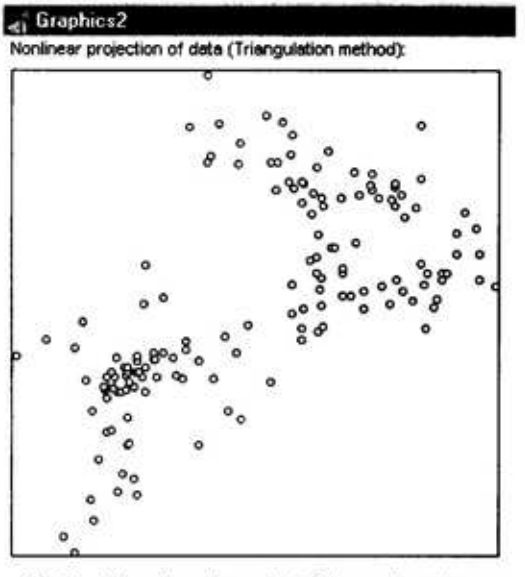
Empaic

File Edit Visualize Hierarchy Partitions Rules Window Help

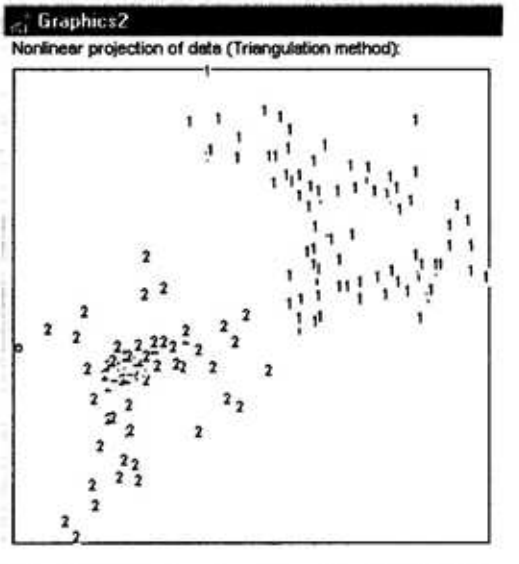
d:\ep\opu\cipp\mitkovic\mag\emp32\cars.ls

	MPG	cylinders	displace	horsepowe	accel	year	weight	origin	maka	model	price
1.	36.1	4	90	48	21.5	78	1985	Europe	Volkswage	Rabbit-DI	2400
2.	36.1	4	98	66	14.4	78	1800	USA	Ford	Fiesta	1900
3.	32.8	4	78	52	19.4	78	1985	Japan	Mazda	GLC-Delux	2200
4.	39.4	4	85	70	18.6	78	2070	Japan	Datsun	B210-GX	2725
5.	36.1	4	91	60	16.4	78	1800	Japan	Honda	Civic-CVC(	2250
6.	19.9	8	260	110	15.5	78	3365	USA	Oldsmobile	Cutlass	3300
7.	19.4	8	318	140	13.2	78	3735	USA	Dodge	Diplomat	3125
8.	20.2	8	302	139	12.8	78	3570	USA	Mercury	Monarch	2650
9.	19.2	6	231	105	19.2	78	3535	USA	Pontiac	Phoenix	2800
10.	20.5	6	200	95	18.2	78	3155	USA	Chevrolet	Malibu	3275
11.	20.2	6	200	85	15.8	78	2965	USA	Ford	Fairmont-A	2375

Sl. 1 – Skup neklasifikovanih primera za učenje



Sl. 2 – Vizuelizacija neklasifikovanih primera



Sl. 3 – Vizuelizacija primera grupisanih u dve klase

```
Best Ruleset (100.00% accuracy):
--------------------------------

Rule  1 (1.000):  [origin=USA] -> [Class=#1]  85
Rule  2 (1.000):  [origin=Europe,Japan] -> [Class=#2]  70
```

Sl. 4 – Naučena pravila za podelu na dve klase

Dakle, naučeno objašnjenje je da se automobili mogu podeliti na #1: automobile američke proizvodnje i #2: ostale automobile, odnosno da se američki automobili bitno razlikuju od svih ostalih. Novi pojmovi još nemaju imena i jednostavno se označavaju kao #1 i #2.

Na taj način može se potražiti objašnjenje i neke druge podele primera, na 2, 3 ili više grupa ili klasa. Koliko je razumno ispitivati? Odgovor daje hijerarhijsko grupisanje, koje generiše kompletnu hijerarhiju razumnih grupisanja primera u hijerarhiju klasa, što je prikazano u desnoj koloni izveštaja algoritma na slici 6.

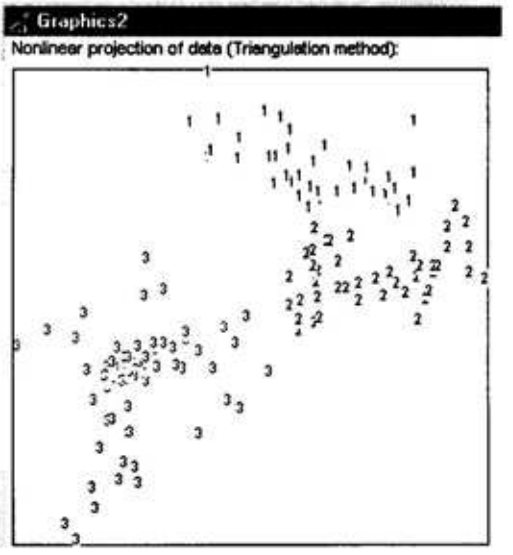
```
if origin=USA
   then Class=#1   (1.0)
   else
if origin=Europe or origin=Japan
      then Class=#2   (1.0)
```

Sl. 5 – Naučena pravila odlučivanja za dve klase

```
-- MNV disaggregative clustering --

MNV treshold    Number of CLASSES
------------    -----------------
40                      2
39                      2
...                     ...
18                      2
17                      3
16                      4
15                      4
14                      6
13                      8
12                      9
11                      9
10                     11
 9                     14
 8                     19
 7                     21
 6                     30
 5                     39
 4                     57
 3                     81
 2                    113
```

Sl. 6 – Izveštaj algoritma hijerarhijskog grupisanja



Sl. 7 – Vizuelizacija nove klasifikacije u tri klase

Induktivno učenje propozicionih koncepata objašnjava samo podelu na disjunktne klase u jednom nivou. Zbog toga se, pre poziva algoritma učenja pravila, prvo izabere odgovarajući nivo hijerarhije, odnosno nivo podele na klase (uokvireno na slici 6). Na slici 7 prikazana je vizuelizacija grupisanja u tri klase:

Rezultat induktivnog učenja za tri klase daje objašnjenje kao na slici 8.

Dakle, u ovoj podeli uočena je podklasa #1 američkih automobila sa 6 i više cilindara (pravilo Rule 1), a #2 su ostali američki automobili raznovrsnih svojstava i jednostavno su opisani nabrajanjem modela (pravilo Rule 2). Evropski i japanski automobili su i dalje kompaktna grupa #3 ostalih automobila (pravilo Rule 3).

Ovaj jednostavan primer ne može u potpunosti da prikaže korist koja se može ostvariti metodama induktivnog učenja, ali su to uočile kompanije i finansijske institucije u razvijenom svetu (npr., banke, osiguravajuća društva, fondovi, trgo-

```
Best Ruleset (100.00% accuracy):
--------------------------------

Rule  1 (1.000):  [displace=>173][cylinders=>6] -> [Class=#1]  41
Rule  2 (1.000):  [model=Fiesta,Phoenix,Fairmont-M,Concord,Chevett
                  Fairmont-4,Colt-Hatch,Spirit-DL,Horizon,HorizonTC3,
                  Must-Cobra,Reliant,Skylark,Aries-SW,Champ,Horizon-4
                  Cavalier,CavalierSW,Cavalier2D,1200-Hatch,Aries-SE,
                  MustangGL,Rampage,Ranger,S-10][cylinders=4] -> [Cls
Rule  3 (1.000):  [origin=Europe,Japan] -> [Class=#3]  70
```

Sl. 8 – Naučeni skup pravila za tri klase

vinski lanci, itd.), koje sve više traže i koriste rezultate inteligentne analize podataka.

Zaključak

Sistemi za inteligentnu analizu podataka mogu da analiziraju sadržaj velikih baza podataka i ustanove određene zakonitosti u tim podacima. Ovako dobijeno znanje može se upotrebiti za donošenje odluka zasnovanih na znanju, npr. u sistemima za dijagnostiku oboljenja ili tehničkih sistema.

Inteligentna analiza podataka je sledeća tehnologija za koju se očekuje da će uskoro početi rutinski da se primenjuje u oblasti informacionog inženjerstva.

U radu je prikazan mali deo bogate ponude komercijalnih alata za ovu namenu. Uočeno je da svi komercijalni sistemi za inteligentnu analizu podataka imaju ugrađenu neku od metoda za induktivno učenje simboličkih opisa, obično stabala odlučivanja ili produkcionih pravila. Opisani (nekomercijalni) sopstveni sistem Empiric takođe koristi navedene metode.

Poređenjem sa rezultatima iz literature zaključeno je da je podsistem za induktivno učenje pravila sistema Empiric po tačnosti predviđanja uporediv sa svetskim standardima u ovoj oblasti.

Literatura:

[1] Michalski, R. S., Kaufman, A.: Data Mining and Knowledge Discovery: A Review of Issues and Multistrategy Approach, in Michalski, R. S., Bratko, I. and Kubat, M.(eds), Machine Learning and Data Mining: Methods and Applications, John Wiley & Sons, 1997.

[2] Mitchel, T. M.: Machine Learning and Data Mining, Communications of the ACM, Vol. 42, No. 11, November 1999.

[3] Goebel, M., Gruenvald, L.: A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, Vol. 1, Issue 1, June 1999.

[4] Michalski, R., Carbonell, J., Miachell, T. (Eds.): Michine learning: An artificial intelligence approach (Vol, I). San Francisco, CA: Morgan Kaufmann, 1983.

[5] Hart, A.: Machine induction as a form of knowledge acquisition in knowledge engineering, in Forsyth, R.(ed), Machine Learning: Principles and techniques, Chapman and Hall, London, 1989.

[6] Graetinger, T.: Diging up $$$ with Data Mining - An Executives Guide, Discovery Corps, Inc., 1999.

[7] Schaffer, C.: A Conservation Law for Generalization Performance, in Proceedings of the Twelfth International Conference on Machine Learning, pp. 259–265, New Brunswick, NJ: Morgan Kaufmann, 1994.

[8] Murphy, P. M. and Aha, D. W.: UCI Repository of machine learning databases [Machine-readable data repository]. Irvine, CA: University of California, Department of Information and Computer Science, 1994.

[9] Kononenko, I. and Šimec, E.: Induction of decision trees using RELIEFF, in Kruse R., Viertl R., Della Riccia G.(eds): CISM Lecture Notes, Springer Verlag, 1995.

[10] Domingos, P.: Unifying Instance-Based and Rule-Based Induction, Machine Learning, 24, pp. 141–168, 1996.

[11] Quinlan, J. R.: Improved Use of Continuous Attributes in C4.5, Journal of Artificial Intelligence Research, Volume 4, pp. 77–90, 1996.

[12] Kohavi, R., Sommerfield, D., Dougherty, J.: Data Mining using MLC ++: A Machine Learning Library in C++, in Tools With AI 1996, pp. 234–245, 1996.

[13] Clark, P., Niblett, T.: The CN2 induction algorithm, Machine Learning, 3, pp. 261–284, 1989.