

Rezime:

U radu se razmatra modeliranje induktivnog mašinskog učenja (inductive machine learning). Izloženi su najznačajniji matematički modeli i njihov odnos prema praktično realizovanim algoritmima. Prikazane su osnovne karakteristike najpoznatijih svetskih i jednog sopstvenog sistema induktivnog učenja, kao i primeri upotrebe sistema u zadacima predviđanja.

Ključne reči: induktivno mašinsko učenje, matematičko modeliranje, model PAC, algoritmi.

MATHEMATICAL MODELING OF INDUCTIVE MACHINE LEARNING

Summary:

This paper considers mathematical modeling of inductive machine learning. Most important mathematical models and their relations to practical realizations of algorithms are reviewed. The basic characteristics of some best known systems and one author's system for inductive rules learning are described and some examples of using systems in prediction tasks are presented.

Key words: inductive machine learning, mathematical modeling, PAC model, algorithms.

Uvod

Indukcija predstavlja izvođenje opštih zakonitosti na osnovu uvida u konkretne pojave, odnosno slučajeve. Problem sa takvim zakonitostima jeste da se njihova istinitost ne može formalno dokazati, već samo opovrgnuti, što su uočili još antički mislioci.

U novije doba F. Bacon je kritikovao nauku XVII veka, smatrajući da predstavlja samo uređen sistem postojećeg znanja, bez usmerenja ka novim otkrićima i metodama otkrivanja. Takođe,

uočio je tendenciju ljudskog mišljenja ka zanemarivanju negativnih primera, i ukazao na opasnost od preterane generalizacije u induktivnom zaključivanju.

Smatrao je da se proces razumevanja mora „izvršiti kao pomoću mašine“ [1]. Prve korake ka tom cilju načinio je J. S. Mill, koji je smatrao da je zadatak logike da obezbedi modele i pravila induktivnog zaključivanja, i predložio svoja četiri pravila indukcije.

Neki savremeni filozofi, kao K. Popper, negiraju postojanje mehanizma poput indukcije u živom svetu i smatraju

da se znanje dobija stvaranjem pretpostavki i odbacivanjem onih koje nisu saglasne sa podacima, odnosno kao pogađanje i eliminisanje greške. Većina sličnih razmatranja odnosi se na proces otkrivanja naučnih i univerzalnih istina, dok se zanemaruje da je veliki deo ljudskog znanja i iskustva koje se koristi u svakodnevnom životu fragmentaran i neprecizan, često sa unapred poznatim brojem kontraprimera.

S druge strane, filozof Bertrand Russell je smatrao da, ako se ne veruje u indukciju, ne veruje se ni u šta. Smatrao je da je induktivni princip u suštini probabilistički i skrenuo pažnju na značaj statističkog zaključivanja.

Važan aspekt modeliranja učenja istakao je i filozof Ludwig Wittgenstein, koji je ukazivao na važnost jednostavnosti, i smatrao da je zadatak indukcije da prihvati kao istinitu najjednostavniju postojeću zakonitost, pošto postoji mnogo generalizacija konzistentnih sa iskustvom (primer opravdanosti ovog stava predstavljaju Ptolomejeva i Kopernikova kosmologija).

Nastankom računara pojavio se i problem konstrukcije mehanizma koji bi mogao da kreira opšte zakonitosti na osnovu konkretnih primera, a kao jednako važan i problem ocene valjanosti tih zakonitosti u praktičnim situacijama.

Inženjerska definicija učenja jeste da je to proces u kojem „sistem poboljšava svoje performanse na datom zadatku bez dodatnog programiranja“ [1].

Mašinski generisano znanje najčešće je namenjeno čoveku i tada je neophodno da bude predstavljeno u obliku koji odgovara ljudskim predstavama i

mentalnim modelima, odnosno da zadovoljava „princip razumljivosti“ (comprehensibility principle) [2]. Takođe, poželjno je da se omogući ne samo predstavljanje i objašnjavanje rezultata, nego i važnijih principa, pretpostavki i teorija koje vode takvim rezultatima.

Jedna od savremenih primena induktivnog učenja jeste u inteligentnoj analizi podataka (data mining, knowledge data discovery) [3]. Osim tradicionalnih metoda, kao što su regresiona analiza, klaster analiza, numerička taksonomija i stohastičko modeliranje, koriste se različite metode mašinskog učenja: fuzzy logika, neuronske mreže, sistemi za induktivno učenje produkcionih pravila i stabala odlučivanja, itd.

Najčešći zadaci inteligentne analize podataka, koji se realizuju tehnikama induktivnog učenja, jesu:

- klasifikacija (classification) – razvrstavanje novih podataka u zadati skup mogućih kategorija;

- predviđanje (prediction) – upotreba prediktivnog modela znanja, tako da se za novi podatak predvidi vrednost zadanog atributa;

- regresija (regression) – izvođenje modela na osnovu analize veza nekih vrednosti atributa sa vrednostima drugih atributa jednog istog podatka. Taj model se zatim može koristiti za predviđanje vrednosti tih atributa u novim podacima.

Modeliranje induktivnog učenja

Formulisanjem matematičkih modela koji omogućavaju rigoroznu formalnu analizu i smisleno poređenje, heuristikom mašinskog učenja bavi se teorija

mašinskog učenja (computational learning theory). Postoji više modela mašinskog učenja [4], [5]:

- model učenja do granice (E. M. Gold, 1967),

- model učenja PAC (L. Valiant, 1984),

- model egzaktnog učenja (D. Angluin, 1988),

- modeli zasnovani na metodama minimalnog kodiranja (Wallace, 1987; Rissanen, 1987),

- modeli učenja na osnovu Bayesove i klasične statističke teorije (W. L. Buntine, 1992),

- ostali pristupi modeliranju induktivnog učenja.

U ovom radu su ovi modeli ukratko opisani, a zatim je detaljnije prikazan najpopularniji model induktivnog učenja – model PAC (Probably Approximately Correct Learning model).

Model učenja do granice

Jednu od najstarijih teorija induktivnog zaključivanja razvio je M. Gold 1967. godine na osnovu teorije rekurzivnih funkcija, odnosno teorije izračunljivosti [6]. Učenje se posmatra kao prikupljanje informacija o nepoznatom objektu, njihovo procesiranje i dobijanje opisa nepoznatog objekta. U idealnom slučaju, to je potpuni opis objekta.

Objekti u ovom modelu su izračunljive rekurzivne funkcije (ili rekurzivno prebrojivi jezici) pomoću kojih se može predstaviti proizvoljno algoritamsko ponašanje, odnosno algoritam učenja. Nepoznata funkcija f zadaje se preko svojih vrednosti $f: f(0), f(1), f(2), \dots$, a algoritam

učenja generiše program h u univerzalnom programskom jeziku¹ koji izračunava nepoznatu funkciju f .

U ovom modelu podrazumeva se raspoloživost neograničenog broja primera za učenje. Broj hipoteza koje sistem može da postavi i ispravi takode nije ograničen, a opis se dobija nakon dovoljnog broja primera, pa se ovaj model učenja naziva i učenje do granice (learning in the limit) [6].

Model učenja PAC

Dok se teorija učenja do granice bavi eventualnom identifikacijom potpuno istinite hipoteze, teorija učenja L. Valianta ima za cilj otkrivanje razumno tačne hipoteze za razumno vreme [6].

Naziv „učenje PAC“ nastao je od namene modela, u kojem uspešno učenje nepoznatog koncepta povlači nalaženje, sa visokom pouzdanošću, hipoteze koja predstavlja njegovu dobru aproksimaciju.

Model PAC pokušava da formalizuje opšti zadatak učenja koncepta, tako da to bude praktično izvedivo (feasible) učenje, koje podrazumeva polinomsku zavisnost vremena izvršavanja i potrebnog broja primera da se koncept nauči. Ovaj model ne zahteva nalaženje potpuno preciznog opisa nepoznatog koncepta, već samo traži da greška aproksimacije bude po volji mala i to sa određenom verovatnošću. Pri tome se podrazumeva da se primeri slučajno uzimaju iz prostora primera u skladu sa distribucijom verovatnoće. Greška aproksimacije izražava se kroz verovatnošću da neki primer pri-

¹ Odnosno, programski jezik dovoljno univerzalan da se u njemu implementira Turingova mašina.

pada skupu primera za koje se ciljni i naučeni koncept razlikuju.

Postoji više proširenja modela PAC radi uključivanja realnih problema u analizu: agnostički (agnostic) model PAC ne pretpostavlja da postoji traženi koncept, dok model PAC sa šumom (noise model) pretpostavlja prisustvo velike količine slučajnog šuma ili malog dela namernog šuma u podacima.

Nedostatak modela PAC mašinskog učenja je egzistencijalna priroda rezultata koje daje: mogućnost ili nemogućnost učenja određenih klasa koncepata, koji malo utiču na osnovni problem konstruisanja programa koji uči.

Model egzaktnog učenja

Model egzaktnog učenja, ili model upita, pred algoritam učenja postavlja zahtev da proizvede hipotezu koja je identična ciljnoj, odnosno klasifikuje sve primere na identičan način kao i nepoznati ciljni koncept. Algoritam učenja može da koristi dodatne izvore informacija, osim obučavajućeg skupa, u obliku upita o klasifikaciji pojedinih primera, upita o pripadnosti i upita o ekvivalentnosti, odnosno provere da li je naučeni koncept već jednak ciljnom.

Jedan od prvih rezultata ovog modela je algoritam D. Angluin za učenje monotonih DNF (disjunctive normal form) formula pomoću upita o pripadnosti i ek-

vivalentnosti. Može se modifikovati za učenje koncepata korišćenjem samo upita o pripadnosti, ako se reše dodatni računski problemi za koje su poznati efikasni algoritmi, što je povezano sa praktičnim problemima na području inteligentne analize podataka (problem otkrivanja asocijacija, nalaženja svih ključeva u relaciji i sl.) [5].

Modeli minimalnog kodiranja

Učenje koncepata može se posmatrati kao kompresija informacija, pošto naučeni koncept (teorija ili hipoteza) opisuje (describes) ili objašnjava (explains) veliki broj primera iz obučavajućeg skupa. U opštem slučaju ne postoji jedinstvena hipoteza koja objašnjava obučavajući skup primera.

Princip MDL (Minimum Description Length) tvrdi da je za objašnjenje podataka najbolja ona teorija koja minimizira sumu dužine opisa same teorije i dužinu podataka zapisanih uz pomoć te teorije (u bitima) [7].

Na slici 1 (a) prikazan je primer podataka koji se sastoje od sekvence dužine 100 000 bita, a jedan program koji opisuje ovu sekvencu na slici 1 (b) (u programu se doslovno nabraja svaki bit). Pošto u sekvenci postoji pravilnost, ovaj program može se skratiti na svega nekoliko bita, kao na slici 1 (c). U ulaznom nizu nema izuzetaka od ove pravilnosti, pa je

```
(a) 00001000010000100001000010000100001 ... 000010000100001  
(b) print '00001000010000100001000010000100001 ... 000010000100001'; halt  
(c) for i=1 to 20000; print '00001'; halt
```

Sl. 1 - Dva načina kodiranja podataka u univerzalnom programskom jeziku

to istovremeno tačan i potpuni opis posmatranog skupa podataka.

Jedan od poznatih principa u izboru optimalne hipoteze ili opisa koncepta je tzv. princip „Okamove oštrice“ (Occam's razor principle), koji nalaže izbor najjednostavnije hipoteze, odnosno one koja minimizira informaciju potrebnu za njen zapis:

$$\arg \min_{h \in H} I(h)$$

gde je $H = \{h_i\}$ skup hipoteza, a $I(h)$ količina informacije potrebna za kodiranje hipoteze $h \in H$. Doslovna primena principa izbora najjednostavnije hipoteze generiše suviše specijalizovane teorije, kojima se opisuju i brojna odstupanja podataka u primerima od tražene zakonitosti.

U prisustvu šuma, određena nekonzistentnost i nekompletnost opisa koncepta je neophodna [8]. Princip MDL ocenjuje ovakve hipoteze i bira one koje minimiziraju količinu informacije potrebne za zapis same hipoteze i zapis obučavajućeg skupa primera pomoću te hipoteze:

$$\arg \min_{h \in H} [I(h) + I(S_i|h)]$$

gde je S_i obučavajući skup primera relativan za hipotezu, a $I(S_i|h)$ je količina informacije potrebna za kodiranje obučavajućeg skupa primera S_i pomoću hipoteze h . Ovaj član predstavlja količinu informacija potrebnih za predstavljanje izuzetaka od teorije, pošto su primeri koje teorija korektno klasifikuje već sadržani u opisu same teorije.

U odsustvu šuma, princip MDL se za nekonzistentne hipoteze svodi na princip

„Okamove oštrice“, jer za njih važi $I(S_i|h) = \infty$, pa se za (kraće) konzistentne hipoteze generiše isti poredak kao i pomoću izraza $I(h)$.

U praksi se pokazalo da, za mali broj primera, mera MDL favorizuje suviše generalne koncepte i ima loše performanse na novim primerima. Ograničenja opisnog jezika, takode, mogu imati kao posledicu nemogućnost generisanja opisa koji dobro objašnjava obučavajuće primere.

Princip MDL je u [9] kombinovan sa tačnošću na obučavajućem skupu u programu za učenje predikatnih, kao i propozicionih koncepta:

$$\arg \min_{h \in H: I(h) + I(S_i|h) < I(S_i)} I(S_i|h)$$

gde je $I(S_i)$ količina informacija potrebna za kodiranje obučavajućeg skupa S_i . Za propozicione koncepte dobijena je ista tačnost predviđanja, uz znatno smanjenje veličine opisa.

Modeli učenja na osnovu Bayesove i klasične statističke teorije

Prema [4], modeli Bayesove statističke teorije predstavljaju dovoljno snažan teoretski okvir za modeliranje svih bitnih aspekata problema mašinskog učenja konceptata, kao što su izračunljivost (kompleksnost) algoritama učenja, ocena greške učenja, predubeđenje, pretraživanje prostora hipoteza, problem detaljisanja (overfitting) i prisustvo „šuma“ u primerima.

Takode, pokazano je da za mnoge probleme koji se razmatraju drugim teorijama mašinskog učenja postoji dualna interpretacija u okviru Bayesove statističke teorije [1], [4].

Ostali pristupi modeliranju induktivnog učenja

U [4] se kao teorije induktivnog učenja navode još i modeli učenja iz klasične matematičke statistike i prepoznavanja uzoraka: diskriminantna analiza, estimacija osnovne gustine (kernel density estimation), metoda najbližeg suseda (nearest neighbour), selekcija i ekstrakcija atributa (feature selection and extraction), kao i metode razvijanja (partitioning methods), koje se u ovom radu neće detaljnije izlagati.

Osnovni rezultati modela induktivnog učenja PAC na osnovu primera

Prema [10] model mašinskog učenja PAC uveo je 1984. godine L. G. Valiant, radi proučavanja računске složenosti algoritama mašinskog učenja, pri čemu je koristio pojednostavljene pojmove iz statističkog prepoznavanja uzoraka i teorije odlučivanja, kao i teorije izračunljivosti, odnosno algoritamske kompleksnosti. Proučavao je izračunljive klase problema mašinskog učenja, odnosno one za koje postoje algoritmi polinomske složenosti koji ih rešavaju.

U originalnom modelu PAC, prostor primera, kao i prostor hipoteza, odnosno koncepata je $\{0,1\}^n$, odnosno sastoji se od niza vrednosti biranih (boolean) atributa. Smisao aproksimacije se definiše uz pretpostavku da postoji neka distribucija D , definisana nad prostorom primera $\{0,1\}^n$, koja svakom primeru pridružuje neku verovatnoću.

Greška hipoteze h , kada je poznat traženi koncept f , definiše se kao verovatnoća da se hipoteza h i koncept f razlikuju na primeru slučajno izabranom u skladu sa distribucijom D :

$$\varepsilon_h = \sum_{x \in \mathcal{X}} D(x)$$

gde Δ označava simetričnu diferenciju.

Hipoteza h je dobra aproksimacija nepoznatog koncepta f ako je greška ε_h dovoljno mala. Najjednostavniji način dobijanja dobre hipoteze jeste traženje nezavisnih slučajnih primera traženog koncepta f , od kojih se svaki sastoji od primera koji je izabran slučajno u skladu sa D i označava se kao pozitivan ako je primer koncepta f , odnosno kao negativan ako nije primer koncepta. Učenje i testiranje koriste istu distribuciju primera, odnosno ne razmatra se „šum“ u podacima. Algoritam učenja je tada izračunljiva procedura koja uzima uzorak traženog koncepta f , koji se sastoji od niza nezavisnih slučajnih primera koncepta f i vraća kao rezultat hipotezu.

Za svaki $n \geq 1$ neka je C_n skup traženih koncepata nad prostorom primera $\{0,1\}^n$ i neka je $\mathbf{C} = \{C_n\}_{n \geq 1}$. Neka su H_n , za $n \geq 1$ i \mathbf{H} definisani na sličan način. Mogućnost učenja PAC (po modelu PAC learnability) može se definisati na sledeći način:

Klasa koncepata \mathbf{C} može se naučiti po modelu PAC u prostoru hipoteza \mathbf{H} ako postoji algoritam učenja koji se izvršava za polinomsko vreme i polinom $p(\cdot, \cdot)$, koji je takav da za sve $n \geq 1$, sve tražene koncepte $c \in C_n$, sve distribucije verovatnoća D nad prostorom primera $\{0,1\}^n$ i svaki ε i δ , gde su $0 < \varepsilon, \delta < 1$, ako

se algoritmu A zadaje najmanje $p(n, 1/\varepsilon, 1/\delta)$ nezavisnih slučajnih primera koncepta c uzetih na osnovu distribucije D , onda algoritam A sa verovatnoćom od najmanje $1/\delta$, vraća hipotezu $h \in H_n$ sa greškom $\varepsilon_n \leq \varepsilon$. Polinom p najmanjeg reda naziva se kompleksnost uzorka (sample complexity) algoritma učenja A .

Definicija ističe da algoritam učenja mora biti efikasan (polinomske složenosti) i mora sa velikom verovatnoćom da daje dobru aproksimaciju nepoznatog koncepta, uz razuman broj slučajno izabranih obučavajućih primera.

Ako se prostor hipoteza H poklapa sa klasom konceptata C , klasu konceptata moguće je pravilno (properly) naučiti prema modelu PAC. Ovaj zahtev nije neophodan kada se traži samo efikasan način predviđanja novih primera, bez obzira na oblik naučene hipoteze.

Neki od poznatijih rezultata teorije PAC o mogućnosti pravilnog učenja po modelu PAC su:

- konjuktivni koncepti mogu se pravilno naučiti prema modelu PAC. Klasa konceptata u obliku disjunkcije dva konjuktivna izraza se ne može pravilno naučiti po modelu PAC, kao ni klasa egzistencijalnih konjuktivnih konceptata u strukturnim prostorima primera sa dva objekta;

- pojedinačni koncepti u obliku perceptrona, odnosno linearnih threshold funkcija, mogu se pravilno naučiti po modelu PAC. Klasa konceptata u obliku konjukcije ili disjunkcije dva koncepta u obliku perceptrona ne može se pravilno naučiti po modelu PAC, kao ni neuralne mreže u više nivoa sa dve sakrivene jedinice;

- klase disjunktivnih i konjuktivnih Boolovih formula konačne dužine k , kao i liste odlučivanja konačne dužine, mogu se pravilno naučiti po modelu PAC za

svako fiksirano k . Nije poznato da li se klase svih disjunktivnih i konjuktivnih Boolovih formula i svih lista odlučivanja mogu pravilno naučiti po modelu PAC.

Dokaz mogućnosti učenja PAC podrazumeva:

- dokazivanje da postoji efikasan algoritam koji nalazi hipotezu u konkretnom prostoru hipoteza i koji je konzistentan (tačan za sve primere) sa zadatim uzorkom bilo kog koncepta iz ciljne klase koncepta,

- dokazivanje da je kompleksnost uzorka svakog takvog algoritma polinomska.

Za veće prostore hipoteza potreban je veći uzorak m primera radi nalaženja konzistentne hipoteze sa zadatom verovatnoćom bar $1/\delta$. Ako se koriste samo diskretni atributi, gornja granica može se iskazati preko kardinalnosti prostora hipoteza $|H_n|$ (tzv. Blumerova granica):

$$m \leq \frac{1}{\varepsilon} \left(\ln |H_n| + \ln \frac{1}{\delta} \right)$$

Kompletan prostor hipoteza nad n propozicionih simbola ima kardinalnost $|H_n| = 2^{2^n}$ (2^n primera, a svaki se može označiti kao istinit ili lažan), a prostor od k disjunkcija konjuktivnih izraza proizvoljne dužine ima kardinalnost $|H_n| \leq (3^n)^k$. Osim kardinalnosti diskretnih prostora hipoteza, za prostore hipoteza sa kontinualnim atributima razmatra se i tzv. VC-dimenzija (Vapnik-Chevo-nenkis dimension). Označava se sa $VC\text{-dim}(H)$ i definiše kao maksimalni broj d primera koji se mogu označiti kao pozitivni i negativni primeri na svih 2^d mogućih načina, tako da je svako označavanje konzistentno sa nekim hipotezama u H .

Pomoću ove veličine može se oceniti gornja granica kompleksnosti, odnosno najmanja potrebna veličina uzorka. Za $C_n \subseteq H_n$ i $n \geq 1$ može se pokazati da svaki konzistentan algoritam za učenje klase C pomoću prostora hipoteza H ima gornju granicu kompleksnosti uzorka:

$$m \leq \frac{1}{\varepsilon(1-\sqrt{\varepsilon})} \left(2 \cdot VC \dim(H_n) \ln \frac{6}{\varepsilon} + \ln \frac{2}{\delta} \right)$$

Model PAC učenja može se jednostavno proširiti na primere sa drugim tipovima atributa, uključujući attribute koji opisuju strukturalna svojstva primera (npr. za opis scene). Moguće je definisati primere kao nizove nad fiksnim alfabetom i ispitivati mogućnost učenja konačnih automata, beskontekstnih gramatika i sličnih klasa koncepata.

Važna modifikacija modela PAC je slabi (weak) model PAC. Pretpostavka tzv. jakog (strong) modela PAC jeste mogućnost postizanja po volji male klasifikacije greške, uz dovoljan broj primera. Slabi model PAC učenja postavlja slabiji zahtev pred algoritam učenja, jer je dovoljno da klasifikacijska greška koju garantuje algoritam učenja bude samo nešto manja od greške slučajne klasifikacije.

Zahtev jakog modela PAC za malom klasifikacijskom greškom može se postići i kombinacijom slabog algoritma PAC (koji zadovoljava uslove slabog modela PAC i posebnog algoritma kombinovanja, kojim se kombinuju rezultati više aktiviranja slabog algoritma nad različitim uzorcima obučavajućih primera. Najpoznatije metode kombinovanja su isticanje (boosting [11]) i natrpavanje (bagging [12]).

Prema [5] jedan od najpoznatijih primera algoritma isticanja, AdaBoost [11], ugrađen je u poznati algoritam učenja C5.0 i sistem za istraživanje podataka Clementine.

Praktično realizovani uspešni sistemi mašinskog učenja, kao što su C4.5, CART i Backprop, ne zadovoljavaju kriterijume modela učenja PAC [5]. Na taj način, PAC model učenja ne omogućava smisleno poređenje različitih algoritama učenja.

Praktično realizovani algoritmi induktivnog učenja

Sistemi induktivnog mašinskog učenja koncepata postoje kao samostalni programi i kao delovi složenijih sistema, kao što su programska okruženja za projektovanje ekspertskih sistema i sistemi za inteligentnu analizu podataka. Prema složenosti koncepata koje su u stanju da nauče, odnosno izražajnoj snazi jezika za opis koncepata, uglavnom se dele na sisteme za propoziciono učenje (predikatni račun nultog reda) i sisteme za predikatno učenje (predikatni račun prvog i, izuzetno, viših redova). Ovakvi koncepti opisuju se izrazima sa rekurzijom i omogućavaju jednostavniji opis složenih objekata.

Najpoznatiji programi za propoziciono učenje su programi iz AQ-familije (Michalski), ID3 (Quinlan), Asisstant-86 (Cestnik, Kononenko, Bratko), C4.5 (Quinlan), CART (Breiman) i CN2 (Clark, Niblett).

Neki od poznatijih programa za predikatno učenje, odnosno induktivno učenje logičkih koncepata, jesu programi tzv. in-

duktivnog logičkog programiranja (inductive logic programming, ILP). Poznatiji sistemi iz ove kategorije su: FOIL (Quinlan), FOCL (Pazzani) i HYDRA (Pazzani).

Autor ovog rada razvio je jedan sistem za predikatno učenje [13, 14] u okviru sopstvenog sistema za inteligentnu analizu podataka Empiric. Prvi prototip sistema autor je razvio na Katedri za računarsku tehniku Visokih vojnotehničkih škola KoV JNA u Zagrebu 1989–1991. godine (verzija za DOS).

Nova multimodelska verzija sistema za induktivno učenje Empiric.Rules razvijena je kao praktični deo izrade magistarske teze na Elektrotehničkom fakultetu u Beogradu [14]. Mogućnost učenja višestrukih modela realizovana je metodom „natrpavanja“ (bagging).

Primeri upotrebe sistema induktivnog učenja

Metodologija upotrebe metoda mašinskog učenja u rešavanju problema predviđanja podrazumeva sledeće korake [15]:

1. Formulisanje problema na način pogodan za tretiranje metodama induktivnog učenja pravila, odnosno u problem predikcije ili klasifikacije;

2. Izbor načina predstavljanja obučavajućih primera i koncepata (izbor i definicija atributa);

3. Prikupljanje podataka za formiranje dobrog skupa primera za učenje;

4. INDUKTIVNO UČENJE (pomoću nekog sistema za automatizovano učenje);

5. Evaluacija naučenog znanja; estimacija tačnosti predviđanja i pregled koji obavljaju stručnjaci za razmatranu problematiku;

6. Primena baze znanja (u okviru nekog sistema zasnovanog na znanju ili nezavisno).

Uspeh, pre svega, zavisi od prve tačke, odnosno od dobre formulacije problema i izbora atributa koji su relevantni za problem. Prednost induktivnog učenja je u tome što je kompleksan zadatak prikupljanja znanja sveden na dva jednostavnija zadatka – definisanje problema i formiranje adekvatne reprezentacije.

Za ilustraciju biće navedeno nekoliko primera primene ove metodologije za rešavanje stvarnih problema u nekoliko oblasti:

- medicinska dijagnostika, problemi Hepatitis, Heart Disease, Thyroid Disease;
- dijagnostika tehničkih sistema, problem Shuttle;
- kriminologija, problem Glass.

Tačka 1. metodologije realizuje se kroz definisanje svakog od navedenih problema [16]:

– *Hepatitis* – cilj je predviđanje preživljavanja pacijenata obolelih od hepatitisa, na osnovu 19 opštih podataka i kliničkih nalaza (age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver-big, liver-firm, spleen-palpable, spiders, ascites, varices, bilirubin, alk-phosphate, sgot, albumin, protime, histology). Neki podaci u primerima su ispušteni;

– *Heart Disease* – zadatak je utvrđivanje da li kod pacijenta postoje znakovi oboljenja srca ili ne, na osnovu 13 pokazatelja (age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar > 120 mg/dl, resting EKG results, max. heart rate, exercise induced angina, ST depression induc by exercise, slope of peak exercise ST segm, # of maj vessel clrd by flourscp, thal);

– *Thyroid Disease* – cilj je dijagnostika oboljenja štitne žlezde: da li pacijent ima oboljenje i koji oblik (originalna varijanta problema R. Quinlan-a), na osnovu 29 opštih podataka i kliničkih nalaza (age, sex, on thyroxine, query on thyroxine, on antithyroid medication, sick, pregnant, thyroid_surgery, I131 treatment, query hypothyroid, query hyperthyroid, lithium, goitre, tumor, hypopituitary, psych, TSH_measured, TSH, T3_measured, T3, TT4_measured, TT4, T4U_measured, T4U, FTI_measured, FTI, TBG_measured, TBG, referral source). Neki podaci u primerima dijagnoza su ispušteni;

– *Shuttle* – cilj je predviđanje stanja radijatora Space Shuttla (7 vrednosti, odnosno klasa), na osnovu numeričkih pokazatelja koji se očitavaju sa senzora (prvi podatak je vreme očitavanja), koristeći znanje u obliku pravila koja može da shvati čovek-operator. Obučavajući skup ima 43 500 klasifikovanih primera očitavanja snimljenih na jednom letu, a testni skup 14 500 primera sa drugog leta;

– *Glass* – cilj je identifikacija vrste stakla pronadenog na licu mesta za kriminološke potrebe, na osnovu indeksa prelamanja (refractive-index) i hemijske analize sastojaka (sodium, magnesium, aluminum, silicon, potassium, calcium, barium, iron). Ispravna identifikacija vr-

ste (porekla) stakla može se upotrebiti kao sudski dokaz.

Tačka 2. metodologije realizovana je tako da je za sve probleme upotrebljen isti (propozicioni) opisni jezik, sa brojem i vrstama atributa, kao u tabeli.

Tačka 3. (prikupljanje podataka) realizovana je u različitim institucijama, npr. za problem Hepatitis na univerzitetu Carnegie-Mellon, a za Shuttle u NASA. Detaljan opis i prikupljeni podaci nalaze se u svetskoj bazi podataka problema induktivnog učenja [16], na adresi www.ics.uci.edu/~mleam/MLRepository.html.

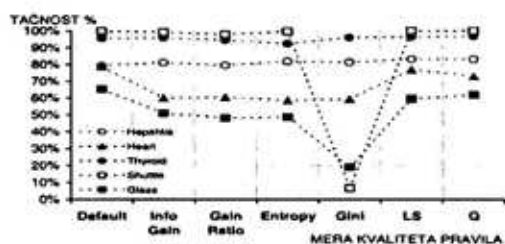
Tačka 4. (induktivno učenje) realizuje se tako da se istovremeno pripreme i osnovni podaci za estimaciju tačnosti iz tačke 5. Deo primera koristi se za učenje, a deo za proveru naučenog znanja, odnosno estimaciju tačnosti predviđanja, slika 2. Skoro objektivna estimacija buduće tačnosti predviđanja dobija se metodom unakrsne validacije (skup primera se na slučajan način podeli na k delova, a zatim se k puta izvrši algoritam učenja, tako da se svaki put $(k-1)/k$ primera koristi za učenje, a $1/k$ za testiranje (k je obično 10).

Kada je broj primera dovoljno veliki (npr. problem Shuttle), obučavajući skup se podeli na dva dela: deo za učenje i deo za testiranje. Proces estimacije tačnosti završava se u jednom prolazu algoritma učenja.

Pregled svojstava opisa problema

Tabela

R. br.	Problem	Broj primera	Broj atributa		Broj klasa	% većinske klase	Ispušt. vredn.
			Diskretnih	Kontinualnih			
1.	Hepatitis	155	13	6	2	79,35%	da
2.	Heart Disease	270	6	7	2	55,56%	-
3.	Thyroid Disease	3772	22	7	3	95,68%	da
4.	Shuttle 43 500	58 000	-	9	7	78,60%	-
5.	Glass	214	-	9	7	35,51%	-



Sl. 2 – Tačnost predviđanja na pojedinim problemima za raspoložive mere kvaliteta

Tačka 5. predviđa i ekspertizu, odnosno detaljnu analizu rezultata (estimacije tačnosti i naučenih pravila) koju obavljaju stručnjaci za određenu problematiku. Ako je znanje namenjeno ljudima, pravila ne smeju da odudaraju od očekivanja stručnjaka, odnosno postojećih mentalnih modela [17].

Za ilustraciju će se uzeti problem Hepatitis (tabela 1). Stručnjaci su povoljno ocenili generisana pravila u kojima su upotrebljeni atributi ALBUMIN, BILIRUBIN i PROTINE, za razliku od manje važnih, kao što su VARICES, SGOT, SPIDERS ili FATIGUE (od ukupno 19 atributa).

Na slici 3 prikazan je tipičan skup pravila koje je generisao sistem Empiric.

```

Rule 1 (1.000): [PROTINE=29..35][AGE=33..70] -> [Class=DIE] 8
Rule 2 (1.000): [ALBUMIN=<2.8000][ALK-PHOSPHATE=86..165] -> [Class=DIE] 7
Rule 3 (0.308): [PROTINE=38..73][ALK-PHOSPHATE=84..280] -> [Class=DIE] 8
Rule 4 (0.667): [PROTINE=50..73][BILIRUBIN=4.1000..7.6000] -> [Class=DIE] 2
Rule 5 (0.118): [ALBUMIN=2.2000..4.2000] -> [Class=DIE] 11
Rule 6 (0.333): [SGOT=60..182][ALK-PHOSPHATE=127] -> [Class=DIE] 2
Rule 7 (1.000): [SGOT=60][AGE=62] -> [Class=DIE] 1
Rule 8 (0.100): [AGE=51..61] -> [Class=DIE] 2
Rule 9 (1.000): [PROTINE=>51][BILIRUBIN=<1.6000] -> [Class=LIVE] 46
Rule10 (1.000): [ALBUMIN=>3.9000][BILIRUBIN=0.7000..1.6000] -> [Class=LIVE] 28
Rule11 (0.596): [PROTINE=41..78] -> [Class=LIVE] 31
Rule12 (0.542): [PROTINE=23..36] -> [Class=LIVE] 26
Rule13 (0.630): [ALBUMIN=>3.7000] -> [Class=LIVE] 17
Rule14 (0.455): [ALBUMIN=2.7000..3.5000] -> [Class=LIVE] 15
Rule15 (0.250): [SGOT=>24] -> [Class=LIVE] 9
Rule16 (1.000): [AGE=23] -> [Class=LIVE] 1

```

Sl. 3 – Skup pravila za problem Hepatitis generisan u jednom od prolaza algoritma Empiric.Rules

Rules, koristeći meru kvaliteta pravila I-content, na kojoj se vidi da je u osnovi saglasan sa navedenim očekivanjima stručnjaka, odnosno sa postojećim znanjem.

Tačka 6. je primena naučenog znanja u praksi, sa ili bez posredovanja računara, što zavisi od ocene njegove tačnosti, stepena razumljivosti i saglasnosti sa postojećim znanjem [14, 17].

Zaključak

Neki principi modeliranja induktivnog učenja poznati su još od antičkih vremena. Mogućnost automatizacije procesa induktivnog zaključivanja razmatrali su F. Bacon i J. S. Mill, a u savremeno doba formira se matematička disciplina koja ima cilj da modelira proces učenja – teorija mašinskog ili automatizovanog učenja (computational learning theory).

U okviru ove teorije razmatrano je više različitih pristupa modeliranju učenja, od kojih se model PAC (Probably Approximately Correct model) najviše koristi u teoretskim analizama. Izloženi

su osnovni teoretski rezultati ovog pristupa modeliranju induktivnog učenja.

Praktični algoritmi induktivnog učenja ne zadovoljavaju osnovni zahtev modela PAC o garantovanju zadate tačnosti, uz dovoljan broj obučavajućih primera. Zato je uveden „slabi“ (weak) model PAC, po kojem je dovoljno da klasifikacijska greška generisanih pravila bude nešto manja od greške slučajne klasifikacije.

Metodologija upotrebe sistema induktivnog učenja i primeri iz nekoliko interesantnih oblasti (medicinska dijagnostika, dijagnostika tehničkih sistema, kriminologija) daju se uz korišćenje autorovog sistema za induktivno učenje pravila na osnovu primera Empiric.Rules.

Izložene teorije induktivnog učenja koriste se za analizu rešivosti određenih klasa problema induktivnog učenja, dok se praktični sistemi induktivnog učenja i dalje zasnivaju na heurističkim procedurama, kao što je npr. klasa algoritama opisana u [14].

Literatura:

[1] Forsyth, R. (ed): Machine Learning: Principles and techniques, London: Chapman and Hall, 1989.
[2] Michalski, R., Carbonell, J., Mitchell T. (Eds.), Machine learning: An artificial intelligence approach (Vol. I). San Francisco, CA: Morgan Kaufmann, 1983.

[3] Mišković, V.: Primena sistema induktivnog mašinskog učenja u inteligentnoj analizi podataka (šira verzija), Vojnotehnički glasnik, 4-5, 429-438, juli-oktobar 2001.
[4] Buntine, W.: A Theory of Learning Classification Rules, PhD thesis, School of Computing Science, University of Technology, Sydney, November 1992.
[5] Domingo, C.: Concept Learning: Theory, Applications and Related Problems, PhD thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, November 1998.
[6] Ambainis, A.: Probabilistic Inductive Inference: A Survey, 1999.
[7] Li, M. and P. Vitányi, P.: Theories of Learning, in An Introduction to Kolmogorov Complexity and Its Applications, Text and Monographs in Computer Science. Springer-Verlag, 1993.
[8] Kaufman, K. A. and Michalski, R. S.: Learning in an Inconsistent World: Rule Selection in AQ18, Reports of the Machine Learning and Inference Laboratory, MLI 99-2, George Mason University, Fairfax, VA, May, 1999.
[9] Kovačić, M.: Stochastic Inductive Logic Programming, PhD, Faculty of Electrical Engineering and Computer Science, University of Ljubljana, Ljubljana, 1994.
[10] Haussler, D.: Probably approximately correct learning, In Proc. of the 8th National Conference on Artificial Intelligence, pages 1101-1108, Morgan Kaufmann, 1990.
[11] Freund, Y., Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting, in Proc of the Second European Conference on Computational learning Theory, March 1995.
[12] Breiman, L.: Bagging predictors, Technical report No. 421, Department of statistics, University of California, Berkeley, CA 94720, September 1994.
[13] Mišković, V.: Sistem za induktivno učenje propozicionih pravila Empiric.Rules, Zbornik radova INFOFEST 2001, 143-150, Budva, septembar 2001.
[14] Mišković, V.: Jedna klasa algoritama za induktivno učenje. Magistarski rad, Elektrotehnički fakultet, Univerzitet u Beogradu, Beograd, novembar 2001.
[15] Langley P., Simon H. A.: Applications of machine learning and rule induction, Communications of the ACM, Vol. 38, No. 11, pp. 54-64, 1995.
[16] 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, 1998.
[17] Mišković, V.: Princip razumljivosti u induktivnom učenju koncepata, Zbornik radova XLV Konferencije ETRAN-a, Sveska III, 123-126, Bukovička Banja, jun 2001.