UTICAJ UPOTREBE STRUKTURNIH ATRIBUTA NA KOMPLEKSNOST INDUKTIVNO NAUČENIH PROPOZICIONIH KONCEPATA

Vladislav Miškovic, *Uprava za sistem logistike*
Sektora za materijalne resurse Ministarstva odbrane SCG, Beograd, Nemanjina 15

Sadržaj – *U radu se razmatra uticaj upotrebe strukturnih umesto nominalnih atributa u induktivnom učenju propozicionih pravila. Ocena složenosti naučenog skupa pravila, odnosno teorije koja opisuje razmatrane podatke, može se upotrebiti kao ocena njihove razumljivosti, što je važan pokazatelj u primenama induktivnog učenja na realne probleme. Daje se teoretski model estimacije složenosti naučenih pravila i rezultati eksperimenata na nekoliko standardnih problema induktivnog učenja.*

1. INDUKTIVNO UČENJE PROPOZICIONIH KONCEPATA

Induktivno učenje se može definisati kao proces u kome sistem poboljšava svoje performanse na datom zadatku bez dodatnog programiranja. Razlikuje se učenje na osnovu primera ili učenje sa učiteljem (*supervised learning*) i učenje posmatranjem i samostalnim otkrivanjem ili učenje bez učitelja (*unsupervised learning*) [1].

Zadatak induktivnog učenja koncepata je računarska indukcija takvih opisa *koncepata* (kategorija, klasa objekata), koji treba da budu u dovoljnoj meri razumljivi ljudima koji se bave određenom problematikom.

Prema "principu razumljivosti" (*principle of comprehensibility*) [1], opisi koncepata treba semantički i strukturno da odgovaraju onima koje bi proizveo čovek prilikom razmatranja istih entiteta. Pri tome treba podjednako lako izraziti i kvalitativne i kvantitativne koncepte, u prirodnom jeziku ili pomoću vizuelnih oblika, crteža i slika [2], [3].

Veoma često se koncepti opisuju propozicionom logikom, koja ima najmanju izražajnu snagu, ali su algoritmi učenja računski veoma efikasni, što je značajno za realne probleme učenja, kod kojih je broj obučavajućih primera veoma veliki. Osim toga, propozicioni opisi koncepata se lako prevode u prirodni jezik.

Propozicioni model problema se sastoji od atributa kojima se opisuju objekti. Prošireni propozicioni model se zasniva na više tipova atributa, koji se razlikuju po vrsti i organizaciji skupa dozvoljenih vrednosti. Nenumerički atributi mogu imati skup vrednosti koji je neuređen, totalno uređen ili parcijalno uređen (obično hijerarhijski) [2]. Atributi sa hijerarhijski organizovanim domenom omogućavaju da se u propozicionoj formi prikažu strukturna svojstva objekata, pa se često nazivaju *strukturnim*.

Mogućnost upotrebe strukturnih atributa je veoma važna za poslovne primene, npr. u analizi podataka iz oblasti e-trgovine, jer omogućava prikaz različitih klasifikacija proizvoda u poropozicionoj formi [4].

2. KOMPLEKSNOST I RAZUMLJIVOST PROPOZICIONIH KONCEPATA

Najjednostavnija mera razumljivosti propozicionih koncepata je kompleksnost opisa koncepta, koja se može grubo oceniti kao ukupan broj elementarnih ispitivanja (uslova, selektora) u opisu koncepta.

Za složeniju kvantifikaciju kompleksnosti, osim različitih empirijskih mera, koristi se tzv. MDL princip (*Minimum Description Length principle*) [5], [6] po kome je najbolja generalizacija ona koja minimizuje sumu dužine opisa koncepta i dužine opisa zadanih primera kodiranih pomoću opisa koncepta (sve u bitima).

Algoritmi induktivnog učenja koncepata zasnivaju se na iterativnom postupku formiranja teorija-hipoteza različite složenosti, koje opisuju obučavajuće primere sa različitom uspešnošću i zatim izboru najverodostojnije od njih za opis naučenog koncepta.

Izbor se vrši pomoću neke empirijske mere kvaliteta opisa (npr. *LEF*[1], *InfoGain*, *Gini*, *ls-content*, *Q-measure*[2]). Tačnost teorije se ocenjuje pogodnom tehnikoma, npr. unakrsnom validacijom.

Estimacija koja se zasniva na MDL principu daje jedinstvenu meru složenosti opisa i tačnosti predviđanja, na osnovu koje se može izabrati najverodostojnija teorija. Prema MDL principu, teorija koja ima najmanju ukupnu dužinu opisa same teorije i opisa obučavajućih primera uz pomoć te teorije predstavlja i najverovatniju teoriju koja ih opisuje. Za učenje se koriste svi obučavajući primeri, pošto je estimacija tačnosti predviđanja već uključena u MDL meru.

Razumljivost se ne može svesti samo na kompleksnost, odnosno konciznost opisa koncepta, već se moraju uzeti u obzir i druge spoznajne karakteristike čoveka [7].

3. PRIMERI INDUKTIVNOG UČENJA KONCEPATA SA STRUKTURNIM ATRIBUTIMA

Kao ilustracija upotrebe strukturnih atributa može se uzeti mali primer prikazan na Sl. 1 [5].

```
class   {+,-}
color-1 {chromatic, achromatic}
color-2 {primary,non-primary, achromatic}
color-3 {blue,red,green,yellow,violet,orange,
         pink, black, white, gray}
shape-1 {convex, non-convex}
shape-2 {polygon,ellipse, straight_lines,curvy}
shape-3 {triangle,hexagon,square,
         proper_ellipse,circle,cross,star,
         kidney-shaped, crescent}
```

```
+,chromatic,non-primary,yellow,convex,polygon,
square
-,chromatic,primary,green,convex,polygon,
hexagon
-,achromatic,achromatic,white,non-
convex,straight_lines,
cross
-,chromatic,primary,?,convex,ellipse,circle
-,achromatic,achromatic,black,convex,ellipse,
circle
+,chromatic,non-primary,pink,convex,?,?
```

Sl. 1. *Primer opisa atributa i obučavajućeg skupa pomoću nominalnih atributa*

Na Sl. 2 je prikazan isti problem korišćenjem strukturnih atributa (u sintaksi sistema *Empiric* [8]).

```
class {+,-}
color [chromatic
       [primary [blue red green]
        non-primary [yellow violet orange pink]]
       achromatic [black white gray]]
shape [convex
       [polygon [triangle hexagon square]
        ellipse [proper_ellipse circle]]
       non-convex
       [straight_lines [cross star]
        curvy [kidney-shaped crescent]]]
```

```
+,yellow, square
-,green, hexagon
-,white, cross
-,primary, circle
-,black, circle
+,pink,convex
```

Sl. 2. *Primer opisa atributa i obučavajućeg skupa pomoću strukturnih atributa*

Vidi se da je problem opisan manjim brojem atributa (samo dva), pa je opis samih primera kraći.

U Tabeli 1 su prikazani skupovi pravila koje su generisali algoritmi učenja pravila iz poznatog okruženja za inteligentnu analizu podataka *WEKA* [9] i sopstvenog sistema za inteligentnu analizu podataka *Empiric*.

Algoritam	Tačnost (CV)	Tačnost (ob. sk.)	skup pravila
C4.5rules	33,3%	100%	`[color-2=primary] -> - [color-1=chromatic] -> + -> -`
C4.5	66,7%	100%	`[color-2=primary] -> - [color-2=non-primary] -> + [color-1=achromatic] -> -`
Ripper	66,7%	66,7%	`color-2=non-primary] -> + -> -`
Empiric (LS)	100%	100%	`[color-2=non-primary] -> + [color-1=achromatic] -> - [color-2=primary] -> -`
Empiric (Default, Q)	100%	100%	`[color-2=non-primary] -> + [color-2=primary,achromatic] -> -`

Tabela 1. *Naučena pravila samo pomoću nominalnih atributa*

U Tabeli 2 je prikazan skup pravila koje je generisao sistem *Empiric* za obučavajući skup sa strukturnim atributima sa Sl. 2 (sistem *WEKA* ne podržava strukturne atribute).

Algoritam	Tačnost (CV)	Tačnost (ob. sk.)	skup pravila
Empiric (Default, ls,Q)	100%	100%	`[color=non-primary]-> + [color=primary,achromatic]->-`

Tabela 2. *Naučena pravila pomoću strukturnih atributa*

Vidi se da je uz pomoć strukturnih atributa dobijen skup pravila koji odgovara kraćem i jasnijem rezultatu iz Tabele 1 i to konzistentno za sve varijante učenja (*Default*, *LS*, *Q*).

U radu se za realniju estimaciju uticaja efekta primene strukturnih atributa na kompleksnost koncepata koriste dva poznata problema induktivnog učenja, *German-credit* i *Adult* [10]. Svojstva ovih problema data su u Tabeli 3.

#	Problem	# primera	# atributa		# klasa	% več. klase	Nep. vred.
			Diskr.	*Kontin.*			
1	**German credit**	1.000	13	7	2	70,00%	ne
2	**Adult**	48.842	8	6	2	76,07%	da

Tabela 3. *Pregled svojstava upotrebljenih primera*

Problem *German Credit* sastoji se u induktivnom učenju pravila za procenu rizika prilikom odobravanja bankarskih kredita. Problem je korišćen u okviru poznatog evropskog projekta iz oblasti statistike i veštačke inteligencije *Statlog*.

Obučavajući skup ima 1000 primera procene zahteva klijenata koji traže kredit. Zahtevi su opisani pomoću 20 atributa (7 numeričkih, 13 nominalnih), a procena se sastoji u razvrstavanju zahteva u jednu od 2 klase (dobar i loš).

Problem *Adult* (poznat i kao *Census Income*, *StatLog* verzija) sastoji se u induktivnom učenju pravila za predviđanje da li će prihod pojedinačnog poreskog obveznika u SAD preći cenzus od 50.000$ godišnje ili ne, na osnovu raspoloživih podataka o obvezniku.

Obučavajući skup se sastoji od podataka poreskog biroa SAD iz 1996. godine za 48.842 poreska obveznika. Obveznici su opisani sa 6 kontinualnih i 8 diskretnih atributa sa nominalnim vrednostima. Oko 7% podataka je ispušteno.

U originalnoj propozicionoj formi, u opisu ovih problema se ne koriste strukturni atributi. Za ovaj eksperiment su neki od nominalnih atributa u svakom od njih transformisani u strukturne, ali tako da obučavajući skup nije promenjen.

U problemu *German Credit* transformisan je nominalni atribut *personal_status*, tako što je izvršena reorganizacija njegove domene prema Tabeli 4.

attribut	nominalni domen	strukturni domen
personal_status	'male_div/sep' 'female_div/dep/mar' 'male_single' 'male_mar/wid' 'female_single'	[male ['male_div/sep' 'male_single' 'male_mar/wid'] female ['female_div/dep/mar' 'female_single']]

Tabela 4: *Konverzija nominalnog atributa u strukturni za problem German Credit*

Novouvedene vrednosti transformisanog atributa (*male*, *female*) mogu se pojaviti samo u generisanim pravilima, dok obučavajući skup ostaje nepromenjen, jer se u primerima koriste samo vrednosti atributa iz terminalnih čvorova strukturnog domena.

U problemu *Adult* transformisani su nominalni atributi *workclass*, *marital-status i native-country* tako što je izvršena reorganizacija njihove domene prema Tabeli 5.

attribut	nominalni domen	strukturni domen
workclasss	Priv1ate Self-emp-not-inc Self-emp-inc Federal-gov Local-gov State-gov Without-pay Never-worked	[Private Self-emp [Self-emp-not-inc Self-emp-inc] Gov [Federal-gov Local-gov State-gov] Without-pay Never-worked]
marital-status	Married-civ-spouse Divorced Never-married Separated Widowed Married-spouse-absent Married-AF-spouse	[Married [Married-civ-spouse Married-spouse-absent Married-AF-spouse] Divorced Never-married Separated Widowed]
native-country	United-States Cambodia England Puerto-Rico Canada ... Yugoslavia El-Salvador Trinadad&Tobago Peru Hong Holand-Netherlands	[North-America [United-States Canada] Middle-America [El-Salvador ...] South-America ... [Ecuador Columbia Peru] Europe [England Greece Yugoslavia ...] Asia [South ...]]

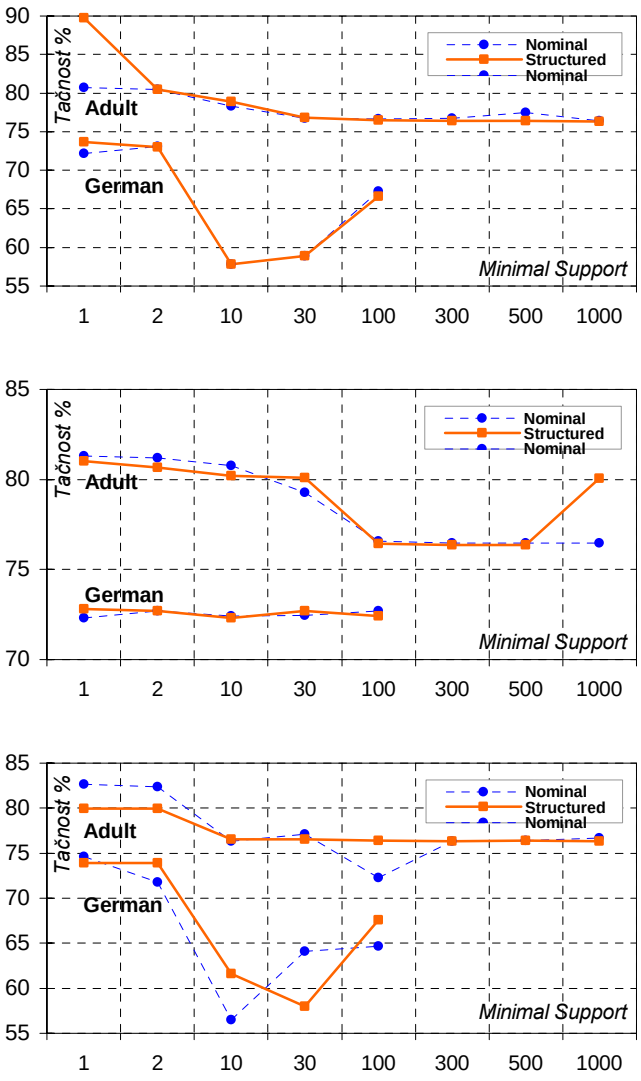
Tabela 5: *Konverzija nominalnih atributa u strukturne za problem Adult*

Estimacija tačnosti predviđanja i kompleksnosti naučenih skupova pravila za oba načina predstavljanja problema izvršena je pomoću sistema *Empiric* [8], a za nominalne atribute rezultati su upoređeni su sa algoritmima induktivnog učenja sistema *WEKA* [9].

4. REZULTATI ESTIMACIJE TAČNOSTI I SLOŽENOSTI NAUČENIH KONCEPATA

Izvršena je estimacija tačnosti i razumljivosti naučenih pravila za tri mere kvaliteta pravila (*Default*, *ls-content* i *Q-measure*) i različite vrednosti parametra *Minimal Support*, koji definiše minimalni broj primera na koje se indukovano pravilo mora oslanjati.

Za problem *German Credit*, tačnost je ocenjena unakrsnom validacijom, a za problem *Adult* metodom *hold-out*, korišćenjem odvojenih skupova primera za učenje i testiranje (32.561 primera za učenje i 16.281 za testiranje). Rezultati estimacije su prikazani na Sl. 3.



Sl. 3: *Tačnost za mere Default, ls-content i Q*

Iz prikaza na Sl. 3 može se uočiti da kod aproksimacije ukupnim brojem selektora, za mere kvaliteta pravila *Default* i *ls-content* upotreba strukturnih atributa ne utiče mnogo na tačnost predviđanja.

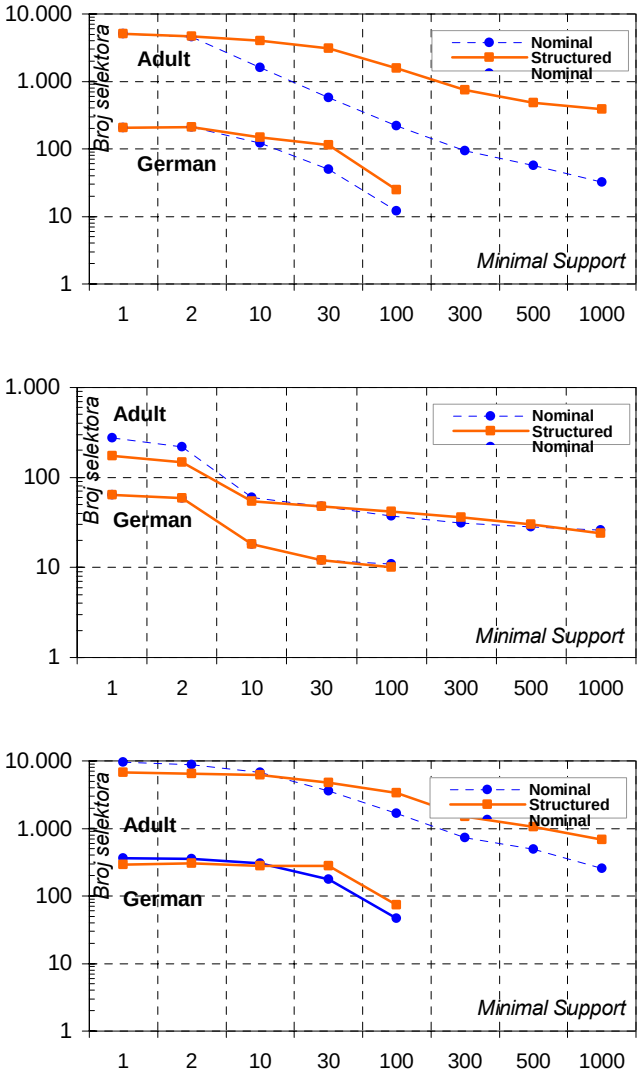
Za meru *Q-measure* tačnost se za skupove sa velikim brojem pravila (vrednosti parametra *Minimal Support* od 1 do 10) za problem *German Credit* povećava, a za problem *Adult* smanjuje.

Za konciznija rešenja (vrednosti parametra *Minimal Support*=100 i više) tačnost predviđanja se upotrebom strukturnih atributa poboljšava.

Estimacija razumljivosti izvršena je preko kompleksnosti naučenih pravila na dva načina:

- pomoću ukupnog broja selektora u skupu pravila [2], [7], Sl. 4 i
- estimacijom MDL-kompleksnosti ([11], [12]), Sl. 5.

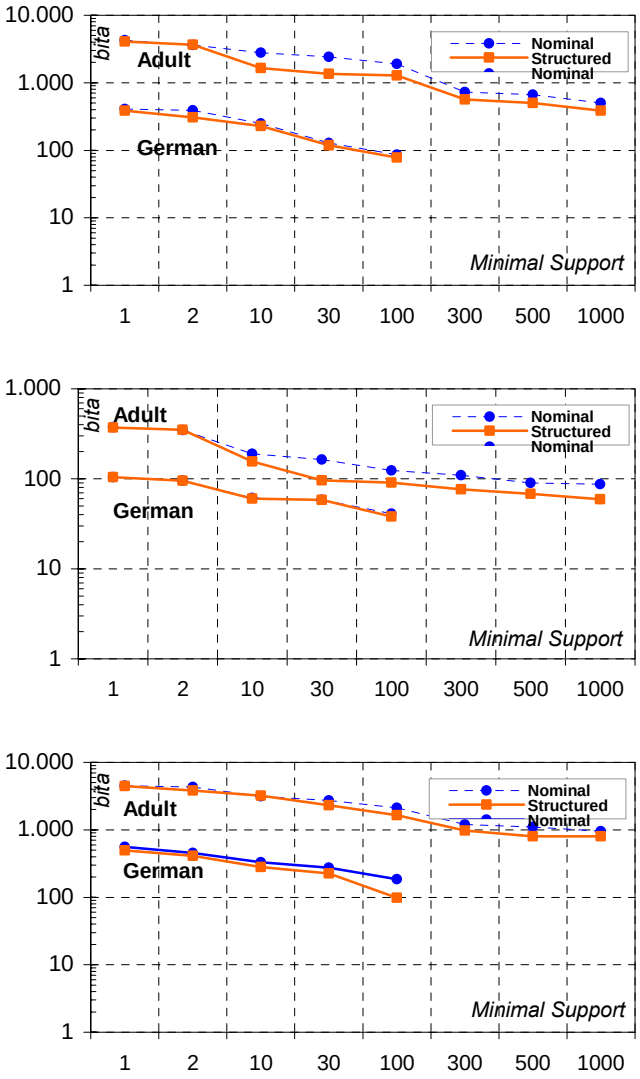
Ocena kompleksnosti sabiranjem broja selektora takođe pokazuje poboljšanje (smanjenje kompleksnosti skupa pravila za strukturne atribute) kod skupova sa većim brojem pravila, odnosno u istim slučajevima kada se povećava tačnost predviđanja, Sl. 4.



Sl. 4: *Kompleksnost za mere Default, ls-content i Q*

Ocena kompleksnosti estimacijom MDL-kompleksnosti pokazuje da se uvođenjem strukturnih atributa za velike skupove pravila kompleksnost povećava, dok se za koncizne skupove pravila, znatnije smanjuje, Sl. 5.

Skupovi pravila manje kompleksnosti su obično razumljiviji, pa se MDL-kompleksnost može koristiti kao ocena razumljivosti u primenama induktivnog učenja na realne probleme.



Sl. 5: *MDL-kompleksnost za mere Default, ls-content i Q*

5. ZAKLJUČAK

Ocena kompleksnosti naučenih pravila jednostavnim sabiranjem broja selektora pokazuje poboljšanja za strukturne atribute u istim slučajevima kada se povećava tačnost predviđanja, koja se meri nekom od standardnih metoda.

Ocena kompleksnosti estimacijom MDL-kompleksnosti pokazuje da se uvođenjem strukturnih atributa za velike skupove pravila kompleksnost povećava, dok se za koncizne skupove pravila znatno smanjuje.

Skupovi pravila manje kompleksnosti su obično razumljiviji, pa se MDL-kompleksnost može koristiti kao ocena razumljivosti u primenama induktivnog učenja na realne probleme.

LITERATURA

[1] Michalski R.S., "A Theory and Methodology of Inductive Learning," in *Michalski, R.S., Mitchell, T. and Carbonell, J. (eds.), Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Co., Palo Alto, pp. 83-134, 1983.

[2] Michalski R.S., "Attributional Calculus: A Logic and Representation Language for Natural Induction", *Reports of the Machine Learning and Inference Laboratory*, MLI 04-2, George Mason University, Fairfax, VA, April, 2004.

[3] 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.

[4] Kohavi R., Mason L., Parekh R., Zheng Z., "Lessons and Challenges from Mining Retail E-Commerce data", *Machine Learning Journal*, 2004.

[5] Cunningham S.J., Humphrey M., Witten I.H., "Understanding what machine learning produces, Part I: representations and their comprehensibility", Working Paper 96/21, Computer Science Department, University of Waikato, 1996.

[6] Conklin D., Witten I.H., "Complexity-Based Induction", *Machine Learning*, Vol.16, No.3, pp. 203-225, 1994.

[7] Pazzani, M., Mani, S. Shankle, W.R., "Comprehensible knowledge-discovery in databases", In *Shafto M.G. and Langley, P. (ed.), Proc. of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 596-601, Lawrence Erlbaum, 1997.

[8] Miškovic V., *Jedna klasa algoritama za induktivno učenje*, Magistarski rad, Elektrotehnički fakultet, Univerzitet u Beogradu, Beograd, novembar 2001.

[9] Witten I. H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, pp. 265-320, Morgan Kaufmann Publishers, 2000.

[10] Blake C.L., Merz C.J., *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[11] Pfahringer B., *Practical Uses of the Minimum Description Length Principle in Inductive Learning*, fur Med. Kybernetik u. AI, Technische Universitat Wien, Dissertation, 1995.

[12] Cleary J.G., Legg S., Witten I.H.., "An MDL Estimate of the Significance of Rules", In *Proceedings of the Information, Statistics and Induction in Science Conference*, pp. 43-53, Melbourne, Australia, 1996.

Abstract – In the paper we consider the effect of using structured instead of nominal attributes in inductive learning of propositional rules. Estimation of rulesets complexity, or complexity of a theory which describes the observed data, can be used as an estimation of their comprehensibility, which is an important landmark in applications of inductive learning to real problems. Theorethical model of estimation and experimental results on several standard inductive learning problems are presented.

THE EFFECT OF USING STRUCTURED ATTRIBUTES TO COMPLEXITY OF INDUCTIVE PROPOSITIONAL CONCEPTS

Vladislav Miškovic