UDK: 628.395/518
Originalni naučni rad

PROCENA KLASE KVALITETA VODA PRIMENOM ALGORITMA NAIVNI BAJES

Zorica SRĐEVIĆ, Bojan SRĐEVIĆ
Univerzitet u Novom Sadu, Poljoprivredni fakultet, Departman za uređenje voda, Novi Sad

REZIME

Cilj rada je da se proveri efikasnost modeliranja i utvrdi mogućnost primene algoritma Naivni Bajes (NB) iz oblasti veštačke inteligencije na proces predikcije klase kvaliteta vode. Za primer je izabrana reka Tamiš i merno mesto Jaša Tomić. U model su uključeni parametri kvaliteta: pH vrednost, suspendovane materije, BPK5, zasićenost kiseonikom i amonijum, a podaci za učenje algoritma su preuzeti iz odgovarajućih baza Agencije za zaštitu životne sredine Republike Srbije za period juli–avgust 2011–2018. Algoritam je tačno predvideo klasu kvaliteta u četrnaest od sedamnaest slučajeva. Sa povećanjem broja parametara kvaliteta i broja slučajeva za učenje, efikasnost i tačnost algoritma bi se dalje povećala što ga preporučuje za korišćenje u proceni kvaliteta voda u različitim vodnim telima, naročito u situacijama kada je važno imati informaciju što pre (na primer, voda za piće ili rekreaciju).

Ključne reči: veštačka inteligencija, Naivni Bajes, klasifikacija, kvalitet voda

1. UVOD

Poslednje decenije svakodnevnog i profesionalnog života ljudi širom sveta karakterišu velike količine informacija i podataka, koje je, da bi bile korisne, potrebno na neki način proveriti, filtrirati, uporediti, klasifikovati itd. Klasični pristupi nisu bili efikasni za ovako kompleksne probleme i razvijeni su novi metodi i tehnike zasnovani na veštačkoj inteligenciji i mašinskom učenju.

Brojni su primeri primene veštačke inteligencije u oblasti vodoprivrede u domaćoj [2, 6, 7, 18] i stranoj literaturi [4, 11, 12, 13, 14, 17, 21]. Najčešće se radi o ekspertskim sistemima, neuralnim i Bajesovim mrežama, evolutivnim tehnikama, fazi logici, raznim heuristikama i meta-heuristikama itd.

Jedan od modela koji se koristi za predikciju u različitim oblastima (medicina, zaštita životne sredine, dubinska analiza podataka - data mining, prepoznavanje lica i govora, biologija, procena rizika, klasifikacija) jeste model Bajesovih mreža (BM) zasnovan na Bajesovoj teoremi za koju je rečeno „da je ona za verovatnoću isto što je Pitagorina teorema za geometriju“[9].

Za razliku od standardne statistike koja na osnovu podataka daje verovatnoće, Bajesova teorema računa naknadne (posteriorne) verovatnoće na osnovu prethodne (aposteriorne) verovatnoće i dostupnih podataka, odnosno „opisuje verovatnoću događaja na osnovu prethodnog znanja o uslovima koji mogu biti vezani za taj događaj“[10].

Bajesove mreže su se pokazale kao efikasne u problemima klasifikacije i zato se danas sve više koriste u tretiranju složenih problema kao što su, na primer, filtriranje spam poruka [3, 5], ili klasifikacija dokumenata [19, 20]; u oba navedena slučaja korišćen je podtip Bajesovih mreža pod nazivom Naivni Bajes.

U ovom radu Naivni Bajes je izabran za formiranje modela za predikciju (procenu) klase kvaliteta voda na osnovu istorije merenja odabranih pet parametara kvaliteta vode u Tamišu, na mernom mestu Jaša Tomić. Pošlo se od toga da je predviđanje kvaliteta vode problem koji je sam po sebi težak, sa jedne strane zbog brojnih faktora koji utiču na kvalitet i od kojih se neki brzo menjaju, a sa druge, zbog posledica za društvo i životnu sredinu. Uzorkovanje, analiza uzoraka, interpretiranje i komuniciranje rezultata zainteresovanim stranama traži vreme koga nekada nema (npr. ako se radi o vodi za piće ili za rekreativne aktivnosti). Razvoj informacionih tehnologija i senzora za praćenje parametara u realnom vremenu su stvorili mogućnost da se do informacije o kvalitetu dođe mnogo brže. Ovde je testirana mogućnost procene

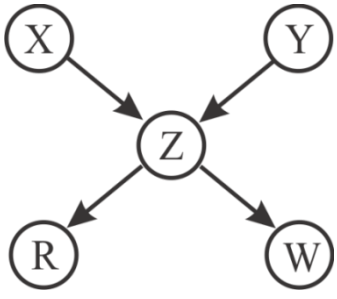
klase kvaliteta vode vodotoka na osnovu učenja iz prethodnih slučajeva stanja kvaliteta vode za pomenuto merno mesto na Tamišu.

Rad je struktuiran na sledeći način. Posle uvodnih razmatranja i napomena, u sledećem poglavlju dati su osnovni pojmovi neophodni za bazično razumevanje Bajesovih mreža, a zatim i matematička formulacija primene Bajesove teoreme na ovakve mreže. Sledi opis algoritma Naivni Bajes i prikaz primera primene za predikciju klasa kvaliteta vode na vojvođanskom vodotoku Tamiš kod naselja Jaša Tomić. U poslednjem poglavlju su dati zaključci i preporuke za dalja istraživanja u predmetnoj oblasti.

2. BAJESOVE MREŽE: OSNOVNI POJMOVI I PRIMENA BAJESOVE TEOREME

Bajesova mreža je, u matematičkom smislu, usmereni aciklični graf *G* koji omogućava efikasno i efektivno predstavljanje zajedničke funkcije raspodele verovatnoće skupa slučajnih promenljivih [8]. Kao takav, graf prikazuje i odnose među promenljivama koje su predstavljene čvorovima i međusobno povezane granama u određenoj strukturi. Primer jednostavne mreže prikazan je na Slici 1.

Čvorovi u mreži mogu imati funkciju: roditelj (X, Y za čvor Z; Z za čvorove R i W), dete (Z za čvorove X i Y; R i W za čvor Z), predak (X, Y za čvorove R i W), potomak (W u odnosu na čvorove X i Y) , nepotomci (svi preci nekog čvora i sam taj čvor), koren (nema roditelje, čvorovi X i Y) i list (nema decu, čvorovi R i W).



Slika 1. Moguća struktura mreže sa pet čvorova

Koreni predstavljaju glavne uzroke problema koji se modelira, a listovi posledice tog problema.

Čvorovi se takođe mogu podeliti na upitne (query) i dokazane (evidence). Putem Bajesove mreže se pomoću vrednosti za dokazane čvorove mogu odrediti verovatnoće za upitne čvorove.

Verovatnoća nad svim promenljivama modelira se pomoću zajedničke funkcije raspodele verovatnoća. Uslovne raspodele verovatnoće su uslovne verovatnoće promenljivih u odnosu na svoje roditelje. Za matematičku formulaciju Bajesove teoreme pogodno je koristiti notaciju iz rada [8].

Neka postoji konačan skup **U** koji čine *n* diskretnih promenljivih: $\mathbf{U} = \{ X_1, X_2, \ldots, X_n\}$. Svaka promenljiva X_i može uzeti vrednost $Val(X_i)$ iz sopstvenog skupa diskretnih vrednosti. Uobičajeno se slučajne promenljive označavaju velikim kurzivnim slovima (npr. *X, Y, Z*), a njihove konkretne vrednosti malim, takođe kurzivnim slovima (npr. *x, y, z*). Skupovi slučajnih promenljivih se obično označavaju velikim boldovanim slovima (npr. **X, Y, Z**). Vrednosti promenljivih u tim skupovima označavaju se malim boldovanim slovima (npr. **x, y, z**); iz očiglednih razloga, analogna notacija onoj gore je $Val(\mathbf{X})$. Konačno, neka je *P* zajednička funkcija raspodele verovatnoća slučajnih promenljivih u skupu **U** i neka su **X**, **Y** i **Z** podskupovi tog skupa. Ako je dato **Z**, kaže se da su **X** i **Y** uslovno nezavisni ako za svako $\mathbf{x} \in Val(\mathbf{X})$, $\mathbf{y} \in Val(\mathbf{Y})$ i $\mathbf{z} \in Val(\mathbf{Z})$ važi da je $P(\mathbf{x}|\mathbf{z},\mathbf{y}) = P(\mathbf{x}|\mathbf{z})$ za $P(\mathbf{y},\mathbf{z}) > 0$.

Bajesova mreža modelira zajedničku raspodelu verovatnoća za skup **U** sa više slučajnih promenljivih. Mreža za skup **U** je par $B = (G, \Theta)$ u kome je *G* oznaka za graf čiji čvorovi reprezentuju slučajne promenljive $X_1, \ldots, X_n$, a linije (kao na Slici 1) predstavljaju direktne zavisnosti između čvorova (slučajnih promenljivih). Graf *G* iskazuje pretpostavku o nezavisnosti, odnosno da je svaka X_i nezavisna od 'potomaka' ako su njeni roditelji u G. Θ predstavlja skup parametara θ koji 'kvantifikuju' mrežu za svaku moguću vrednost x_i od X_i i Π_{x_i} od Π_{X_i} na sledeći način:

$$\theta_{x_i|\prod x_i} = P_B(x_i|\Pi_{x_i}),$$

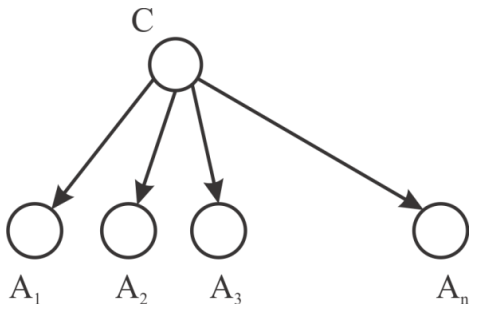
gde je Π_{X_i} skup roditelja od X_i u grafu *G*.

Tada Bajesova mreža *B* definiše jedinstvenu zajedničku raspodelu verovatnoća nad **U** kao:

$$P_B(X_1, X_2, .., X_n) = \\ = \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{x_i}} \quad . \qquad (1)$$

3. ALGORITAM NAIVNI BAJES

Naivni Bajes (NB) se često naziva i NB klasifikator. Spada u najjednostavnije, ali iznenađujuće efikasne modele u okviru Bajesovih mreža. Naziva se 'naivan' jer se zasniva na pretpostavkama da su svi atributi podjednako važni i da su statistički nezavisni. Najčešće se predstavlja kao mreža gde je čvor klasa *C* jedini roditelj čvorova atributa A_i (i=1,..,n) [8]; Slika 2.



Slika 2. Struktura mreže kod primene algoritma Naivni Bajes

Ideja je da algoritam uči iz j skupova slučajeva ('trening podataka'). Neka je **U**(A_1, A_2, …, A_n, C) gde su promenljive $A_1(j)$, $A_2(j)$, …, $A_n(j)$ atributi, a promenljiva $C(j)$ je klasa. C je koren mreže i atributi imaju samo jednog roditelja, tj. $\Pi_C = \emptyset$ i $\Pi_{A_i} = \{C\}$.

Pretpostavka o nezavisnosti atributa bitno pojednostavljuje računanje uslovnih verovatnoća atributa (neki autori koriste i termin 'očekivanje'). Iz definicije uslovne verovatnoće sledi

$$P_r(C|A_1, A_2 \cdots, A_n) = \alpha * P_r(C) * \prod_{i=1}^{n} P_r(A_i|C) \quad (2)$$

gde je α normalizaciona konstanta. Ovo je u stvari definicija Naivnog Bajesa koja se najčešće navodi u literaturi [8], odnosno:

$$P_r(\boldsymbol{A}|C) = P_r(A_1, A_2 \cdots, A_n|C) = P_r(A_1|C) * P_r(A_2|C) * \cdots P_r(A_n|C) * P_r(C)$$

$$P_r(C|\boldsymbol{A}) = = P_r(A_1|C) * P_r(A_2|C) * \cdots P_r(A_n|C) * P_r(C)/P_r(\boldsymbol{A}).$$

Verovatnoće svih kombinacija atributa $A_1(j)$, $A_2(j)$, …, $A_n(j)$ za j-ti slučaj i korespodentni ishod C za taj slučaj definišu predikciona pravila za određivanje klase ishoda. Znajući ova pravila, klasifikacija se vrši računanjem verovatnoće određene klase $C(j+1)$ na osnovu novog, datog skupa podataka $A_1(j+1)$, $A_2(j+1)$, …, $A_n(j+1)$ i predviđanjem klase sa najvećom posteriornom verovatnoćom.

Problem učenja Bajesove mreže se može definisati na sledeći način: Dat je skup trening podataka D = {$\mathbf{u}_1$, $\mathbf{u}_2$, …,$\mathbf{u}_N$}, gde svako $\mathbf{u}_i$ sadrži vrednosti za svaku promenljivu u skupu **U.** Treba naći Bajesovu mrežu *B* koja najbolje odgovara *D*.

Način računanja verovatnoća pripadnosti (j+1)-og skupa atributa određenoj klasi biće prikazan na sledećem jednostavnom primeru. Pretpostavimo da studenti treba da idu na teren da uzorkuju vodu. Odluka da li će se ići na teren (C) ima dva ishoda –'da' ili 'ne'– koji zavisi od tri atributa:

- vremenska prognoza (A_1) - sunčano, oblačno, kiša;
- temperature (A_2) – vruće, hladno, umereno i
- jačina vetra (A_3) – jak, slab.

Trening podaci (**U**) su dati u Tabeli 1.

Ako se zna da je napolju sunčano, postoji 67% šanse da će se ići na teren (u podacima za trening vidi se da se od tri puta kada je bilo sučano, dva puta išlo na teren). Uslovna verovatnoća je:

$$P_r(\text{Teren} = \text{da}|\text{Vremenska prognoza} = \text{sunčano}) = 0{,}67.$$

U tabeli 2 date su učestanosti pojavljivanja za svaki od ishoda (Teren=da, Teren=ne) za sve atribute i za svaki ishod.

Učestanosti pojavljivanja iz Tabele 2 se pretvaraju u uslovne verovatnoće $P_r(\boldsymbol{A}|C)$ i $P(C)$, koje formiraju Tabelu 3 uslovnih verovatnoća.

Ako se zna da su vremenski uslovi

Vremenska prognoza = sunčano (A_1)
Temperatura = umereno (A_2)
Jačina vetra = jak (A_3)

očekivanje da će studenti ići na teren (Teren = da) za date vremenske uslove je

$$P_r(\text{Teren} = \text{da}|\mathbf{A}) = = P_r(A_1|\text{Teren} = \text{da}) * P_r(A_2|\text{Teren} = \text{da}) * P_r(A_3|\text{Teren} = \text{da}) * P_r(\text{Teren} = \text{da}) = 0{,}40 * 0{,}4 * 0{,}2 * 0{,}62 = 0{,}0198.$$

Očekivanje da se pod ovim uslovima ipak neće ići na teren će biti

$$P_r(\text{Teren} = \text{ne}|\mathbf{A}) = = P_r(\text{A}_1|\text{Teren} = \text{ne}) * P_r(\text{A}_2|\text{Teren} = \text{ne}) * P_r(\text{A}_3|\text{Teren} = \text{ne}) * P_r(\text{Teren} = \text{ne}) = 0{,}33 * 0{,}33 * 0{,}67 * 0{,}38 = 0{,}0277.$$

Tabela 1. Trening podaci za problem odlaska na teren

ATRIBUTI			KLASA
Vremenska prognoza (A_1)	Temperatura (A_2)	Jačina vetra (A_3)	Teren (C)
sunčano	vruće	slab	ne
sunčano	vruće	jak	da
oblačno	umereno	slab	da
kiša	hladno	jak	ne
oblačno	vruće	slab	da
sunčano	hladno	slab	da
kiša	umereno	slab	da
oblačno	umereno	jak	ne

Tabela 2. Učestanosti pojavljivanja za problem odlaska na teren

Vremenska prognoza	Teren DA	Teren NE	Temperatura	Teren DA	Teren NE	Jačina vetra	Teren DA	Teren NE	Teren	
sunčano	2	1	vruće	2	1	jak	1	2	da	5
oblačno	2	1	umereno	2	1	slab	4	1	ne	3
kiša	1	1	hladno	1	1					
Ukupno	5	3		5	3		5	3		8

Tabela 3. Uslovne verovatnoće atributa i verovatnoća ishoda

Vremenska prognoza	Teren DA	Teren NE	Temperatura	Teren DA	Teren NE	Jačina vetra	Teren DA	Teren NE	Teren	
sunčano	0,40	0,33	vruće	0,40	0,33	jak	0,20	0,67	da	0,62
oblačno	0,40	0,33	umereno	0,40	0,33	slab	0,80	0,33	ne	0,38
kiša	0,20	0,33	hladno	0,20	0,33					
Ukupno	1,00	1,00		1,00	1,00		1,00	1,00		1,00

Verovatnoća da će pod ovim uslovima studenti ići na teren je

$$P_r(\text{Teren} = \text{da}|\mathbf{A}) = \frac{P_r(\text{Teren} = \text{da}|\mathbf{A})}{P_r(\text{Teren} = \text{da}|\mathbf{A}) + P_r(\text{Teren} = \text{ne}|\mathbf{A})} = \frac{0{,}0198}{0{,}0198 + 0{,}0277} = 0{,}42$$

odnosno 42%.

Da bi se odredila predikciona pravila za određivanje klase ishoda, neophodno je odrediti verovatnoće svih kombinacija atributa i ishoda. Čak i za primer sa malim brojem atributa i ishoda, broj kombinacija je veliki da bi se računalo ručno i očigledno je potrebna softverska podrška.

Za centralni cilj ovog rada – izvršiti procenu klase kvaliteta vode kod mernog mesta Jaša Tomić na Tamišu u Vojvodini – korišćen je softver Netica [15], jedan od najpoznatijih softvera za Bajesove mreže.

4. PROCENA KLASE KVALITETA VODE POMOĆU BAJESOVE MREŽE I ALGORITMA NAIVNI BAJES

4.1 Opis problema i ulazni podaci

Da bi se ocenila tačnost predikcije algoritma Naivni Bajes (NB) i primenljivost na problem klasifikacije kvaliteta voda u vodnom telu, formirana je mreža za procenu klase kvaliteta vode na osnovu učenja iz prethodnih slučajeva za vodotok Tamiš.

Koren mreže je klasa 'Kvalitet vode', a atributi mreže su parametri kvaliteta. Od velikog broja mogućih parametara, za proračune su izabrani oni sa najvećim težinskim vrednostima pri računanju indeksa kvaliteta vode za vodotoke u Srbiji (Serbian Water Quality Index – [16]). Na osnovu težina, datih na sajtu Agencije za zaštitu životne sredine, izabrani su sledeći parametri: pH vrednost, suspendovane materije, BPK5, zasićenost kiseonikom i amonijum (Tabela 4).

Tabela 4. Težine parametara kvaliteta vode: SWQI i stara klasifikacija (adaptirano prema http://www.sepa.gov.rs/index.php?menu=46&id=8012&akcija=showExternal)

	SWQI	Stara klasifikacija parametara kvaliteta			
Parametri	Težina	I klasa	II klasa	III klasa	IV klasa
Zasićenost kiseonikom	18	90-105	75-90 105-115	50-75 115-125	30-50 125-130
BPK5	15	2	4	7	20
Amonijum	12	0,1	0,1	0,5	0,5
pH vrednost	9	6,8-8,5	6,8-8,5	6,0-9,0	6,0-9,0
Ukupni oksidi azota	8	10,05	10,05	15,50	15,50
Ortofosfati	8	0,005	0,01	0,01	0,01
Suspendovane materije	7	10	30	80	100
Temperatura	5	-	-	-	-
Elektroprovodljivost	6	-	-	-	
E. Coli	12	2.000	100.000	200.000	200.000

Podaci za mernu stanicu Jaša Tomić za letnji period (juli, avgust) za 2011 – 2018 godinu, Tabela 5, preuzeti su sa OPEN DATA platforme Agencije za zaštitu životne sredine [1] i godišnjih izveštaja Agencije.

Za svaki od sedamnaest skupova podataka izračunat je indeks kvaliteta vode SWQI, prikazan u poslednjoj koloni Tabele 5.

Tabela 5. Podaci za mernu stanicu Jaša Tomić (juli, avgust 2011 – 2018.) i vrednosti SWQI

Slučaj	ATRIBUTI					KLASA
	pH vrednost (A_1)	Suspendovane materije (A_2)	BPK5 (A_3)	Zasićenost kiseonikom (A_4)	Amonijum (A_5)	SWQI (C)
1	7,8	26	1,5	86	0,05	odlican
2	8,1	20	3	104	0,07	veoma dobar
3	7,9	5	3	118	0,02	odlican
4	8	3	1,1	92	0,06	odlican
5	8,2	4	2	104	0,03	odlican
6	8,21	6	1,2	100	0,03	odlican
7	8,25	5	1,5	102	0,02	odlican
8	7,92	112	1	85	0,1	veoma dobar
9	7,85	60	2,5	77	0,07	dobar
10	8,28	8	1	107	0,02	odlican
11	8,36	4	1	96	0,02	odlican
12	7,6	107	2,8	80	0,07	dobar
13	7,7	102	1,4	76	0,06	dobar
14	8	8	1,4	88	0,02	odlican
15		22	2,2	92	0,06	veoma dobar
16	7,7	72	1,7	89	0,04	veoma dobar
17	7,8	25	1,2	85	0,05	odlican

4.2 Primena NB algoritma

Podaci iz ove tabele predstavljaju podatke za učenje Bajesove mreže. Ima 17 slučajeva, 5 atributa (A_1: pH-vrednosti, …, A_5: Amonijum) i jedan čvor klase (SWQI).

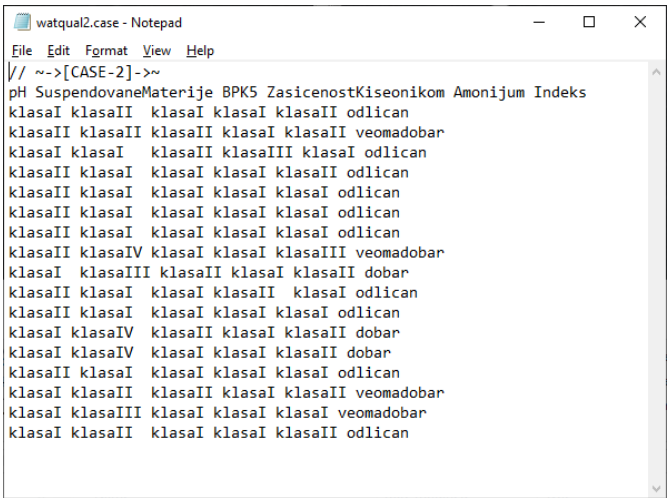
Da bi se primenio NB algoritam, kontinualne veličine je potrebno diskretizovati. Za parametre suspendovane materije, BPK5 i zasićenost kiseonikom diskretizacija je urađena putem klasifikacije date u Tabeli 4; npr. zasićenost kiseonikom od 108% pripada klasi III za ocenu kvaliteta vodotoka.
Svi izmereni podaci za pH vrednost i amonijum pripadali su klasi I tako da je dodatna klasifikacija izvršena na osnovu raspona izmerenih vrednosti; za pH, klasa I obuhvata vrednosti 7,37-7,94, a klasa II 6,8-7,37 i 7,94-8,51. Izmerene vrednosti amonijuma kretale su se od 0,02 do 0,1. Da bi se dobile tri klase, razlika je podeljena na tri intervala. U klasu I svrstane su vrednosti 0,02-0,047, u klasu II vrednosti 0,047-0,074, a u klasu III vrednosti 0,074-0,1.

Sve vrednosti parametara su posle klasifikacije diskretne kao što je prikazano u Tabeli 6.

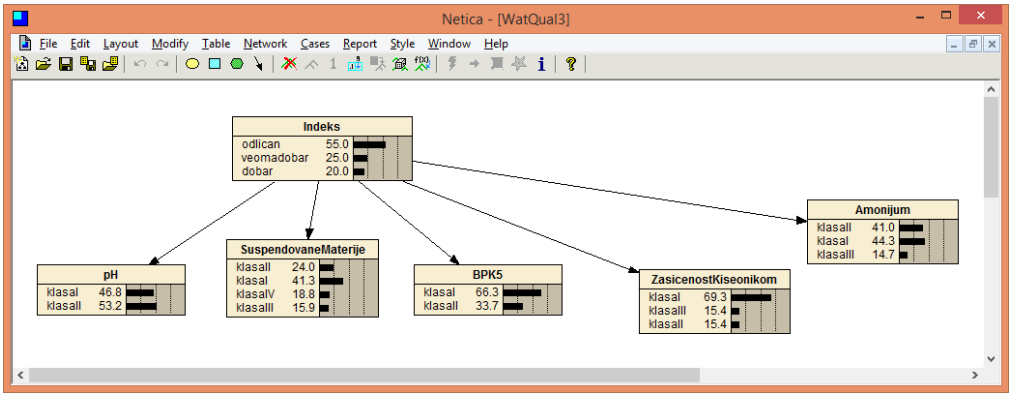
Svaki od sedamnaest redova Tabele 6 predstavlja jedan slučaj u skupu slučajeva za učenje algoritma NB. Tabela 6 (bez poslednje kolone) je uneta u tekst-fajl (Slika 3a) u formi odgovarajućoj za softver Netica, a zatim je fajl importovan u softver da bi se formirao model mreže (Slika 3b) putem opcija Cases-Learn-Add Case file nodes i Cases-Learn-Incorp Case file.

Tabela 6. Klasifikovane vrednosti parametara, SWQI i NB procena klase kvaliteta

	pH vrednost	Suspendovane materije	BPK5	Zasićenost kiseonikom	Amonijum	SWQI	NB procena
1	klasa I	klasa II	klasa I	klasa I	klasa II	odlican	veoma dobar
2	klasa II	klasa II	klasa II	klasa I	klasa II	veoma dobar	veoma dobar
3	klasa I	klasa I	klasa II	klasa III	klasa I	odlican	odlican
4	klasa II	klasa I	klasa I	klasa I	klasa II	odlican	odlican
5	klasa II	klasa I	klasa I	klasa I	klasa I	odlican	odlican
6	klasa II	klasa I	klasa I	klasa I	klasa I	odlican	odlican
7	klasa II	klasa I	klasa I	klasa I	klasa I	odlican	odlican
8	**klasa II**	**klasa IV**	**klasa I**	**klasa I**	**klasa III**	**veoma dobar**	**veoma dobar**
9	klasa I	klasa III	klasa II	klasa I	klasa II	dobar	dobar
10	klasa II	klasa I	klasa I	klasa II	klasa I	odlican	odlican
11	klasa II	klasa I	klasa I	klasa I	klasa I	odlican	odlican
12	klasa I	klasa IV	klasa II	klasa I	klasa II	dobar	veoma dobar
13	klasa I	klasa IV	klasa I	klasa I	klasa II	dobar	dobar
14	klasa II	klasa I	klasa I	klasa I	klasa I	odlican	odlican
15	*	klasa II	klasa II	klasa I	klasa II	veoma dobar	veoma dobar
16	klasa I	klasa III	klasa I	klasa I	klasa I	veoma dobar	odlican
17	klasa I	klasa II	klasa I	klasa I	klasa II	odlican	veoma dobar



Slika 3a. Ulazni fajl za softver Netica sa trening podacima



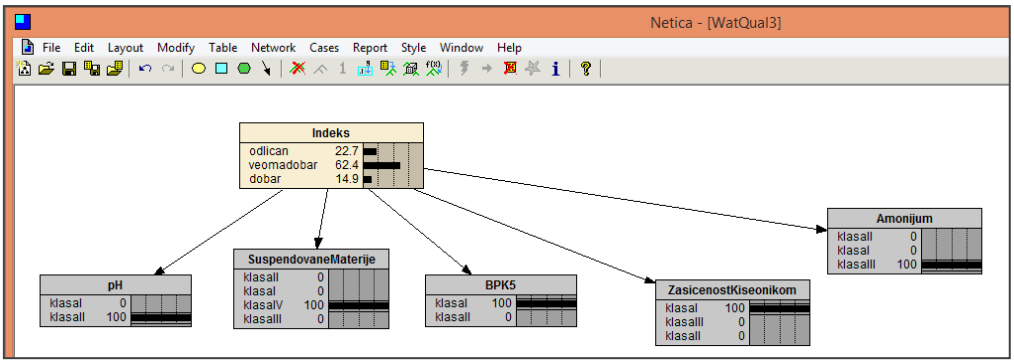
Slika 3b. Bajesova mreža za procenu kvaliteta vode reke Tamiš

Verovatnoće koje su prikazane u svakom čvoru mreže su verovatnoće 'naučene' na osnovu 17 slučajeva.

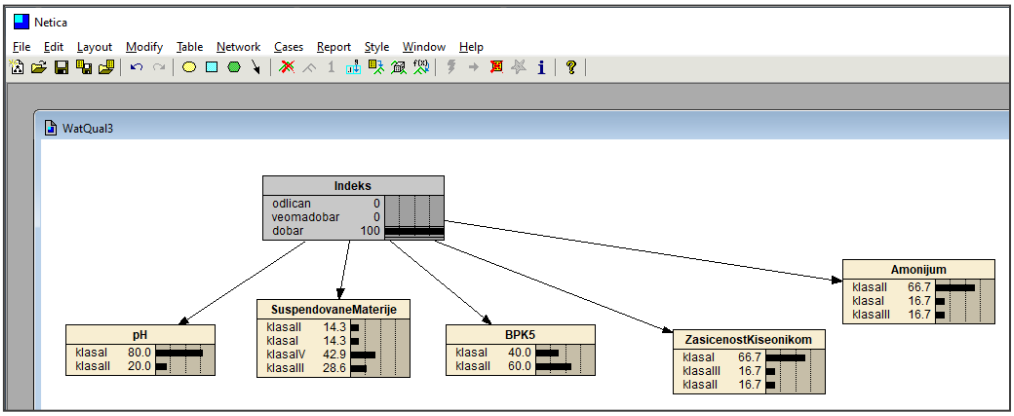
Formirana mreža omogućava da se računaju posteriori i aposteriori verovatnoće indeksa i parametara kvaliteta voda ako se sa sigurnošću znaju klase nekog ili svih parametara, ili se zna indeks kvaliteta uzorka. Na primer, za slučaj pod rednim brojem 8 iz Tabele 6 (pH – klasa II, suspendovane materije – klasa IV, BPK5 – klasa I, zasićenost kiseonikom – klasa I, amonijum – klasa III) izračunato je da je najveća verovatnoća da će uzorak sa takvim karakteristikama pripadati klasi VEOMA DOBAR sa verovatnoćom 64.2% (videti Sliku 4).

Korišćenjem NB mreže za svaki od sedamnaest slučajeva je procenjena najverovatnija klasa kvaliteta vode koja je data u poslednjoj koloni Tabele 6 (NB procena). Od 17 slučajeva, u samo tri slučaja se NB procena ne poklapa sa indeksom koji je izračunat preko sajta Agencije za zaštitu životne sredine.

Sa druge strane, ako se zna da je kvalitet vode ocenjen kao DOBAR, najverovatnija kombinacija parametara koji daje takav indeks je pH – klasa I, suspendovane materije – klasa IV, BPK5 – klasa II, zasićenost kisonikom – klasa I, amonijum – klasa II (videti Sliku 5).



Slika 4. NB procena klase kvaliteta vode na osnovu podataka za učenje NB algoritma



Slika 5. Verovatne klase parametara za poznat kvalitet vode

5. ZAKLJUČAK

Veštačka inteligencija i mašinsko učenje pružaju velike mogućnosti za rešavanje kompleksnih problema, pod uslovom da postoji odgovarajući skup kvalitetnih podataka za trening inteligentnih mreža, kao što su npr. Bajesove. Digitalizacija u oblasti voda, korišćenje senzora i IoT tehnologija ubrzava formiranje baza sa kvalitetnim i obimnim podacima i pruža nove mogućnosti za korišćenje ovih matematičkih alata.

Za očekivati je da veći broj slučajeva za trening rezultira većom tačnošću rezultata bilo koji od inteligentnih modela da se koristi. U radu je prikazan primer predikcije klasa kvaliteta vode reke Tamiš na jednom mernom mestu. Pokazano je da se, ako se koristi Bajesova mreža, i pri relativno malom skupu trening podataka dobijaju rezultati sasvim zadovoljavajuće tačnosti. Uključivanje većeg broja parametara i dužeg niza podataka (više tzv. 'slučajeva') u formiranje modela je sledeći korak u testiranju efikasnosti NB algoritma.

Korišćenje softvera Netica je znatno olakšalo sva potrebna računanja i omogućilo da se, nakon formiranja mreže i učenja mreže iz prethodnih slučajeva, na jednostavan način analiziraju uzročno-posledične veze između izabranih parametara i klasa kvaliteta vode. Preporuka za dalja istraživanja mogućnosti Bajesovih mreža i algoritama, ne samo Naivnog Bajesa, vezana je svakako za vodna tela kod kojih pitanja kvaliteta vode imaju naročit značaj.

ZAHVALNOST

Sredstva za realizaciju istraživanja obezbeđena su od strane Ministarstva za prosvetu, nauku i tehnološki razvoj republike Srbije (ugovor 451-03-68/2020-14/ 200117).

LITERATURA

[1] Agencija za zaštitu životne sredine (2020) OPEN DATA platforma. (http://data.sepa.gov.rs/dataset/kvalitet-voda, pristup 25.03.2020.).

[2] Božić I., Jovanović R. (2018) Standardni i savremeni pristupi u određivanju energetskih karakteristika velikih i malih hidroelektrana. Zbornik Međunarodnog kongresa o procesnoj industriji – Procesing 31: 49-62. Dostupno na: <https://izdanja.smeits.rs/index.php/ptk/article/view/3450>.

[3] Chen J., Dai Z., Duan J. et al. (2019) Improved Naive Bayes with optimal correlation factor for text classification. SN Appl. Sci. 1, 1129.

[4] Chou J., Ho C., Hoang H. (2018) Determining quality of water in reservoir using machine learning. Ecological Informatics 44: 57-75.

[5] Dada E. G., Bassi J. S., Chiroma H., Abdulhamid S. M. et al. (2019) Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5(6), e01802.

[6] Đorđević B. (2019) Smer razvoja hidrotehničke infrastrukture u procesu transformacije naselja u 'pametne' gradove. Vodoprivreda 51: 31-54.

[7] Đorđević B., Dašić T. (2015) Ekspertni sistemi za planiranje i operativno sprovođenje odbrane od poplava. Vodoprivreda 47: 187-202.

[8] Friedman N., Geiger D., Goldszmidt M. (1997) Bayesian network classifiers. Mach. Learn. 29: 131–163.

[9] Harold J. (1998) [1961] The Theory of Probability (3rd ed.). Oxford, England.

[10] Joyce J. (2019) Bayes' Theorem. The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward Zalta (ed.). (https://plato.stanford.edu/archives/spr2019/entries/bayes-theorem/ pristup 10.04.2020.)

[11] Kiiza C., Pan S., Bockelmann-Evans B., Babatunde A. (2020) Predicting pollutant removal in constructed wetlands using artificial neural networks (ANNs). Water Science and Engineering 13(1): 14-23.

[12] Miraki S., Zanganeh S.H., Chapi K. et al. (2019) Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. Water Resour Manage 33: 281–302.

[13] Moazamnia M., Hassanzadeh Y., Nadiri A. A., Khatibi R., Sadeghfam S. (2019) Formulating a strategy to combine artificial intelligence models using Bayesian model averaging to study a distressed aquifer with sparse data availability. Journal of Hydrology 571: 765- 781.

[14] Molina J., Zazo S. (2017) Causal Reasoning for the Analysis of Rivers Runoff Temporal Behavior. Water Resour Manage 31: 4669–4681.

[15] Netica (2020) Norsys Software Corp. (https://www.norsys.com/netica.html, pristup 15.03.2020.)

[16] Serbian Water Quality Index (2020) Agencija za zaštitu životne sredine. (http://www.sepa.gov.rs/index.php?menu=46&id=8012&akcija=showExternal, pristup 30.03.2020.).

[17] Šiljić Tomić A., Antanasijević D., Ristić M., Perić-Grujić A., Pocajt V. (2018) Application of experimental design for the optimization of artificial neural networkbased water quality model: a case study of dissolved oxygen prediction. Environ. Sci. Pollut. Res. 25: 9360–9370.

[18] Stević S., Durović Ž. (2007) Fazi ekspertsko upravljanje izborom optimalnog hidroenergetskog resursa. Vodoprivreda 39: 215-224.

[19] Xu S. (2018) Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science 44(1): 48–59.

[20] Zhang H., Wei H., Tang Y., Pu Q. (2019) Research on Classification of Scientific and Technological Documents Based on Naive Bayes. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 327–331.

[21] Zhao L., Dai T., Qiao Z., Sun P., Hao J., Yang Y. (2020) Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection 133: 169-182.

WATER QUALITY CLASS PREDICTION USING NAIVE BAYES ALGORITHM

by

Zorica SRĐEVIĆ, Bojan SRĐEVIĆ
University of Novi Sad, Faculty of Agriculture, Department of Water Management, Novi Sad, Serbia

Summary

Aim of this paper is to analyse modeling capacity and efficiency in application of the Naïve Bayes (NB) algorithm for predicting water quality classes. River Tamiš in Serbia at Jaša Tomić measuring point is selected as a case study. The algorithm is used to clasify water quality for given water body based on the following parameters: pH, suspended matterial, BOD5, oxygen saturation and ammonium. Data for training algorithm are taken from the database of Environmental Protection Agency of Serbia for the period July–August 2011–2018.

Although relatively simple, the NB classifier correctly predicted water quality class in fourteen out of seventeen cases. It is reasonable to expect increased efficacy and accuracy of this algorithm if more parameters and more cases are available; this recommends its use for water quality classification for different water bodies, especially if information on water quality is needed in real time such as for consuming drinking water or using it for recreational purpose.

Key words: artifical inteligence, Naïve Bayes, classification, water quality

Redigovano 5.11.2020.