МАТФ
Универзитет у Београду
Математички факултет

Metode za rešavanje problema simboličke regresije

master rad

Jana Jovičić 1097/2019

mentor: dr Aleksandar Kartelj

Matematički fakultet
Univerzitet u Beogradu

26. septembar 2022.

Sadržaj

Simbolička regresija (SR)

- Pronalazak matematičkog izraza u simboličkoj formi, koji dobro modeluje vezu između ciljne promenljive i nezavisnih promenljivih.
- Istovremeno uči i o strukturi modela i njegove parametre.
- Problem kombinatorne optimizacije.
- Smatra se da je NP-težak problem.
- Ako je dat skup podataka (X_i, y_i), $i = 1, ..., m$, gde $X_i \in \mathbb{R}^n$ predstavlja i-ti skup atributa, a $y_i \in \mathbb{R}$ i-tu ciljnu promenljivu, cilj SR je pronalazak funkcije $f : \mathbb{R}^n \to \mathbb{R}$ koja najbolje odgovara skupu podataka, odnosno za koju važi $y_i \approx f(X_i), i = 1, ..., m$.
- Izraz se može predstaviti pomoću sintaksnog stabla.

Evaluacija modela simboličke regresije

- Ranije - u terminima metaheuristike kojom je problem rešavan.
- Poslednjih godina - pomoću metrika kao što je koeficijent determinacije R^2 uz podelu skupa podataka na trening i test deo.

$$R^2 = 1 - \frac{MSE}{\mathrm{Var}(y)}$$

- Smatra se da je ciljna funkcija f ispravno određena kandidatskom funkcijom f' ako algebarska simplifikacija izraza $f' - f$ daje simbol "0".

Algoritam grube sile

- Sistematična pretraga prostora matematičkih izraza.
- Pretraga radi iterativno po visini sintaksnog stabla izraza.
- U prvoj iteraciji terminali su samo promenljive. U svakoj narednoj u skup terminala se dodaju i stabla kreirana u prethodnoj iteraciji.
- Postupak se ponavlja sve dok se ne pronađe tačno rešenje (izraz sa MSE manjom od 10^{-6}) ili dok se ne dostigne definisano vremensko ograničenje.
- Ograničenja memorijskih resursa.

Generisanje početne populacije i funkcije prilagođenosti

Generisanje početne populacije:

1. "Full"metoda – Generisanje potpunog stabla.
2. "Grow"metoda – Generisanje stabala čiji oblici variraju.
3. "Ramped half-and-half" metoda – Generisanje stabala različitih visina i oblika.

Funkcije prilagođenosti:

1. "Raw" – zbir distanci između pravih i predviđenih vrednosti izraza.

$$r(i,t) = \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

2. Standardizovana – redefiniše "raw" t.d. bolje jedinke imaju manju vrednost funkcije.

$$s(i,t) = r(i,t)$$

Funkcije prilagođenosti (nastavak)

3. "Adjusted" – dodatno se ističe bolja od dve posmatrane dobre jedinke, veća je za bolje jedinke.

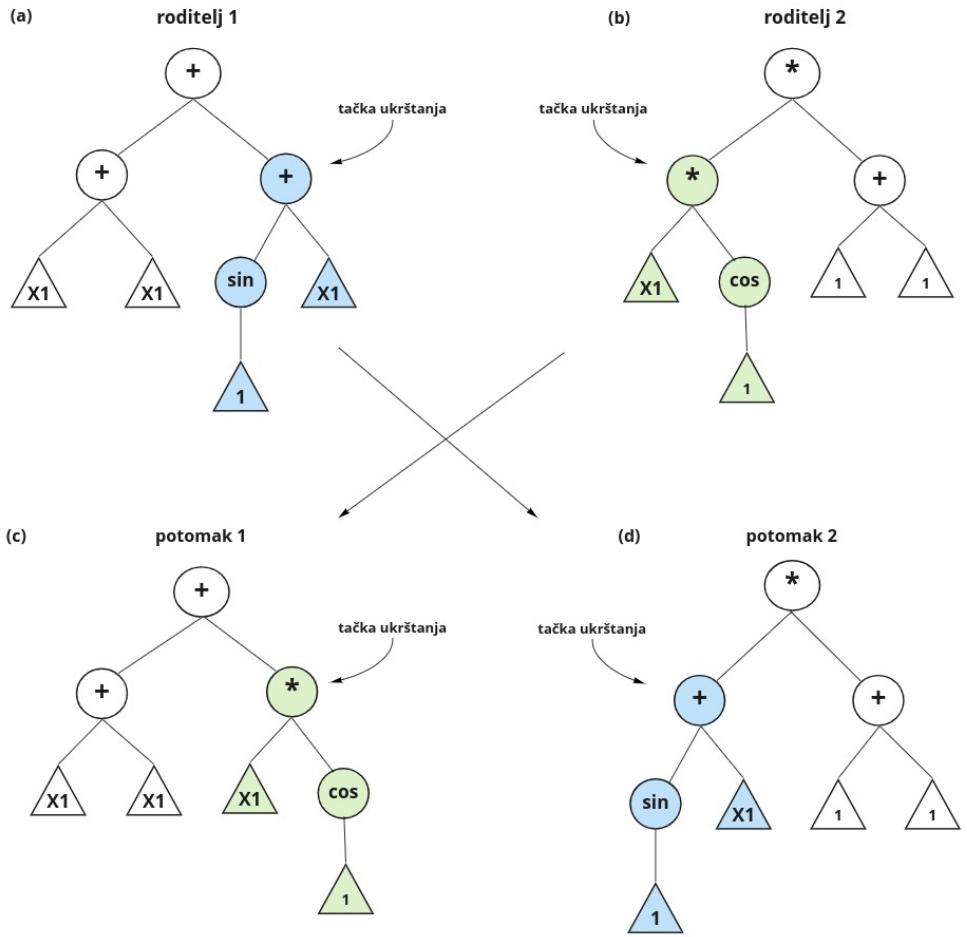
$$a(i,t) = \frac{1}{1+s(i,t)} \in [0,1]$$

Najčešće se koristi. Definiše i kriterijum zaustavljanja – pronalazak jedinke sa "adjusted" funkcijom većom od 0.9.

4. Normalizovana – normalizacija "adjusted" funkcije jedinke u skladu sa "adjusted" funkcijama ostalih jedinki iz populacije.

$$n(i,t) = \frac{a(i,t)}{\sum_{k=1}^{M} a(k,t)},$$

Operatori ukrštanja - Standardni operator ukrštanja



Operatori ukrštanja - Operator zasnovan na semantičkoj sličnosti

- *Semantika uzorkovanja* (SS) nekog podstabla se aproksimira pomoću vrednosti dobijenih evaluacijom tog podstabla na predefinisanom skupu tačaka iz domena problema.
- Neka je F funkcija koja je izražena pomoću (pod)stabla T na domenu D i neka je P skup tačaka iz domena D, $P = \{p_1, p_2, ..., p_N\}$. Tada je *semantika uzorkovanja* stabla T na skupu P u domenu D, skup $S = \{s_1, s_2, ..., s_N\}$ takav da je $s_i = F(p_i), i = 1, 2, .., N$.
- *Rastojanje semantike uzorkovanja* (SSD) između dva podstabla: Neka je $P = \{p_1, p_2, ..., p_N\}$ *semantika uzorkovanja* podstabla St_1, a $Q = \{q_1, q_2, ..., q_N\}$ *semantika uzorkovanja* podstabla St_2. Onda se SSD između St_1 i St_2 definiše kao

$$SSD(St_1, St_2) = \frac{1}{N}(|p_1 - q_1| + |p_2 - q_2| + ... + |p_N - q_N|).$$

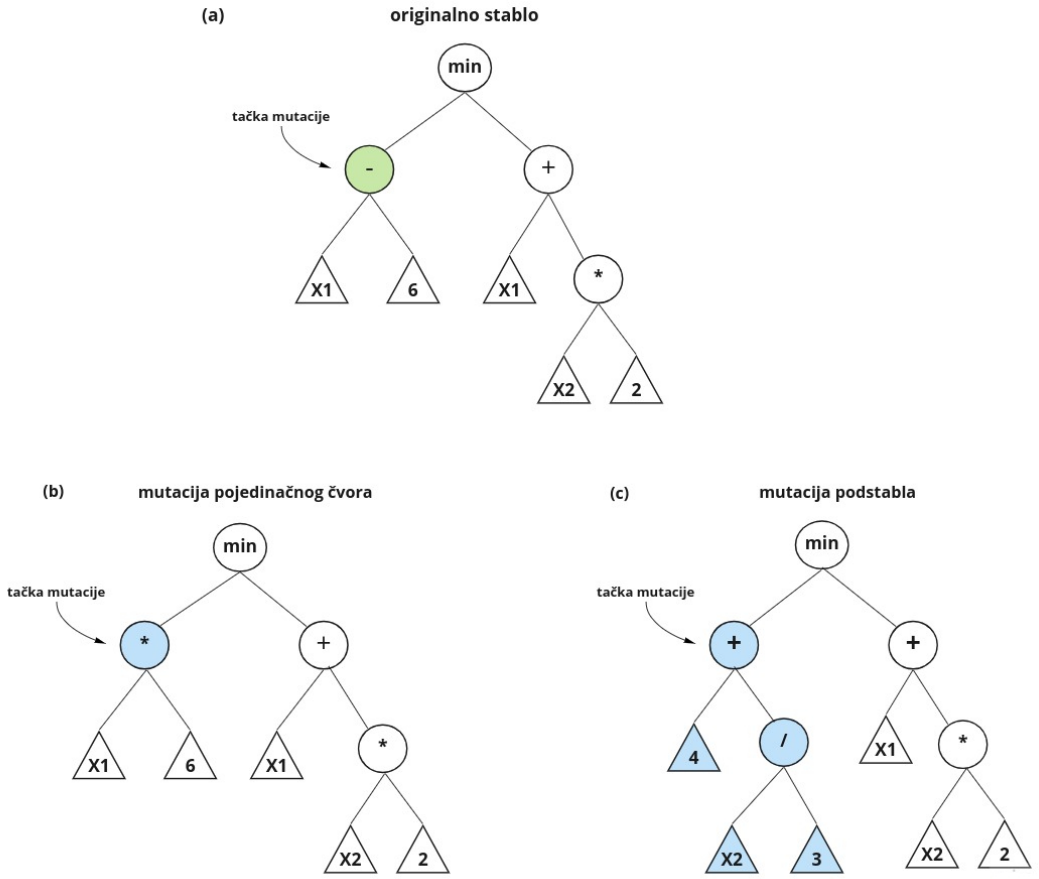
Operatori ukrštanja - Operator zasnovan na semantičkoj sličnosti

- Dva podstabla su *semantički slična* (SS_i) na domenu ako njihova SSD vrednost leži na nekom pozitivnom intervalu.

$$SS_i(St_1, St_2) = \begin{cases} true, & \text{ako je } \alpha < SSD(St_1, St_2) < \beta \\ false, & \text{inače} \end{cases}$$

- Operator ukrštanja zasnovan na semantičkoj sličnosti (SSC) – ukrštanje samo semantički sličnih podstabala.
- Koristi se veći broj pokušaja za pronalazak semantički sličnog para.
- Ako se pređe dozvoljeni broj pokušaja, podstabla se biraju na slučajan način.

Operatori mutacije



Metoda promenljivih okolina

1. Inicijalizacija: Izbor skupa okolina $N_k, k = 1, ..., k_{max}$; Konstruisanje početnog rešenja x;
2. Ponavljanje narednih koraka sve dok se ne ispuni kriterijum zaustavljanja:
   1. Postaviti k = 1;
   2. Ponavljati naredne korake sve dok je $k \leq k_{max}$
      1. *Razmrdavanje* - Generisanje slučajnog rešenja x_1 iz okoline $N_k(x)$;
      2. *Lokalna pretraga* - Primeniti neku metodu lokalne pretrage sa početnim rešenjem x_1. Rezultat pretrage označiti sa x_2;
      3. *Prihvatanje rešenja i promena okoline* - Ako je dobijeno rešenje x_2 bolje od x, postaviti $x = x_2$ i $k = 1$; Inače, postaviti $k = k + 1$;

Kriterijum zaustavljanja može biti pronalazak tačnog rešenja (određen pomoću R^2), maksimalan dozvoljeni broj iteracija ili maksimalno vreme izvršavanja.

Tipovi okolina i razmrdavanje

Tipovi okolina:

1. $N(T)$ - struktura susedstva koje se koristi tokom lokalne pretrage. Članovi ove vrste susedstva se formiraju elementarnim transformacijama stabla.
2. $N_1(T)$ - struktura susedstva čiji se članovi dobijaju mutacijom pojedinačnog čvora.
3. $N_2(T)$ - struktura susedstva čiji se članovi dobijaju mutacijom celog podstabla.

Okoline $N_1(T)$ i $N_2(T)$ se koriste u proceduri razmrdavanja.

Razmrdavanje:

1. Dobija se k-ti sused stabla T, primenom istog poteza k puta.
2. Prvo se nasumičnobira okolina $N_1(T)$ ili $N_2(T)$, a zatim se taj operator primenjuje k puta nad datim stablom.

Elementarne transformacije stabla (ETT) i lokalna pretraga

- Elementarne transformacije stabla (ETT): Neka je $G(V, E)$ neusmereni graf i neka je $T(V, A)$ neko razapinjuće stablo grafa G. ETT transformiše stablo T u stablo T' (u oznaci $T' = ETT(T)$) sledećim koracima:
  - U stablo T dodati granu a, takvu da $a \in E \setminus A$.
  - Detektovati formirani ciklus i ukloniti bilo koju granu (osim one koja je dodata u prethodnom koraku) iz njega kako bi se dobio podgraf T', koji takođe predstavlja razapinjuće stablo grafa G.
- Strategija *prvog poboljšanja*
- Poređenje kvalieta stabala – koeficijent determinacije R^2.

Podaci

Sve metode su testirane pomoću tri vrste skupova podataka:

1. Skup podataka generisan na osnovu funkcija koje se često razmatraju u literaturi. Za svaku funkciju je generisano 100 instanci na osnovu slučajno odabranih vrednosti nezavisnih promenljiih.

$F_1 = x^3 + x^2 + x, \quad x \in [-1, 1]$
$F_2 = x^4 + x^3 + x^2 + x, \quad x \in [-1, 1]$
$F_3 = x^5 + x^4 + x^3 + x^2 + x, \quad x \in [-1, 1]$
$F_4 = x^6 + x^5 + x^4 + x^3 + x^2 + x, \quad x \in [-1, 1]$
$F_5 = sin(x^2)cos(x) - 1, \quad x \in [-1, 1]$
$F_6 = sin(x) + sin(x + x^2), \quad x \in [-1, 1]$
$F_7 = log(x + 1) + log(x^2 + 1), \quad x \in [0, 2]$
$F_8 = sin(x_0) + sin(x_1^2), \quad x_0, x_1 \in [-1, 1]$
$F_9 = 2sin(x_0)cos(x_1), \quad x_0, x_1 \in [-1, 1]$

Podaci

2. Skup podataka generisan na osnovu jednostavnijih funkcija radi upoređivanja metaheurističkih metoda sa metodom grube sile.
   $F_{01} = x_0 x_1 + x_1$
   $F_{02} = x_1 + x_1^2 + x_0$
   $F_{03} = x_0 x_1 + cos(x_0)$
   $F_{04} = x_0 - x_1 x_1$
   $F_{05} = x_0 - x_1 x_1 + x_1$
3. Jedan od javno dostupnih skupova podataka za regresiju - "Yacht Hydrodynamics" skup. Skup sadrži 308 instanci koje su određene pomoću 6 nezavisnih i jedne ciljne promenljive. Sve vrednosti su realnog tipa.

Rezultati

Poređenje sa algoritmom grube sile:

- Algoritam grube sile – optimalno rešenje za $F_{01}, ..., F_{05}$ i F_1 i F_2. U ostalim primerima je dolazilo do nedostatka memorijskih resursa.
- Svi metaheuristički pristupi - optimalno rešenje za primere $F_{01}, .., F_{05}$.

Poređenje metaheurističkih metoda:

- Svaki skup podataka je podeljen na trening (70%) i test (30%) deo.
- Svaka metoda je evaluirana tako što je pokrenuta po 30 puta nad svim skupovima podataka – u svakom pokretanju je dobijen najbolji izraz.
- Za taj izraz je izračunat R^2 na trening i test skupu i provereno je da li i simbolički odgovara ciljnom izrazu.
- Pri svakom pokretanju mereno je i vreme koje je bilo potrebno za pronalazak najboljeg rešenja.

Evaluacija i poređenje metaheurističkih metoda

Tabela: Prosečne vrednosti određenih karakteristika u 30 nezavisnih pokretanja

	Prosečna R^2 vrednost na trening skupu			Prosečna R^2 vrednost na test skupu		
	GP	**GP sa SSC**	**VNP**	**GP**	**GP sa SSC**	**VNP**
F_{01}	0.879	0.831	**0.949**	0.898	0.861	**0.945**
F_{02}	0.765	0.802	**0.848**	0.791	0.818	**0.897**
F_{03}	0.745	0.733	**0.863**	0.810	0.820	**0.926**
F_{04}	0.733	0.740	**0.914**	0.820	0.775	**0.922**
F_{05}	0.795	0.771	**0.846**	0.797	0.765	**0.813**

Evaluacija i poređenje metaheurističkih metoda

Tabela: Prosečne vrednosti određenih karakteristika u 30 nezavisnih pokretanja

	Broj pokretanja u kojima je pronađeno rešenje simbolički ekvivalentno ciljnom rešenju			**Prosečno vreme izvršavanja (s)**		
	GP	**GP sa SSC**	**VNP**	**GP**	**GP sa SSC**	**VNP**
F_{01}	7	3	**13**	12	19	**4**
F_{02}	1	3	**11**	7	13	**6**
F_{03}	5	5	**14**	6	12	**5**
F_{04}	2	1	**9**	12	18	**5**
F_{05}	3	1	**9**	**7**	18	**7**

Evaluacija i poređenje metaheurističkih metoda

Tabela: Prosečne vrednosti određenih karakteristika u 30 nezavisnih pokretanja

	Prosečna R^2 vrednost na trening skupu			Prosečna R^2 vrednost na test skupu		
	GP	GP sa SSC	VNP	GP	GP sa SSC	VNP
F_1	**0.914**	0.861	0.907	**0.907**	0.827	0.872
F_2	**0.827**	0.799	0.771	**0.824**	0.798	0.770
F_3	**0.851**	0.851	0.797	0.695	-1.428	**0.752**
F_4	0.746	0.691	**0.809**	**0.796**	0.743	0.778
F_5	0.643	0.606	**0.894**	0.607	0.589	**0.887**
F_6	0.928	0.917	**0.945**	0.883	0.881	**0.930**
F_7	0.960	0.968	**0.994**	0.950	0.959	**0.993**
F_8	0.857	0.837	**0.968**	0.716	0.657	**0.936**
F_9	0.950	0.940	**0.963**	0.955	0.938	**0.971**
Yacht	0.238	0.213	**0.477**	0.264	0.233	**0.457**

Evaluacija i poređenje metaheurističkih metoda

Tabela: Prosečne vrednosti određenih karakteristika u 30 nezavisnih pokretanja

	Broj pokretanja u kojima je pronađeno rešenje simbolički ekvivalentno ciljnom rešenju			Prosečno vreme izvršavanja (s)		
	GP	**GP sa SSC**	**VNP**	**GP**	**GP sa SSC**	**VNP**
F_1	0	0	**13**	15	24	**7**
F_2	0	0	**9**	13	26	**9**
F_3	0	0	**2**	20	23	**10**
F_4	0	0	**2**	14	27	**12**
F_5	0	0	0	**14**	24	**14**
F_6	0	0	**1**	13	27	**12**
F_7	0	0	0	18	51	**13**
F_8	0	0	**3**	13	35	**7**
F_9	1	1	**2**	12	28	**8**
Yacht	/	/	/	183	247	**30**

Evaluacija i poređenje metaheurističkih metoda

Tabela: Informacije o izrazu koji daje maksimalnu R^2 vrednost na test skupu od svih izraza dobijenih pri 30 nezavisnih pokretanja

	Maksimalna R^2 vrednost na test skupu			Izraz koji ima maksimalnu R^2 vrednost na test skupu		
	GP	GP sa SSC	VNP	GP	GP sa SSC	VNP
F_{01}	1.0	1.0	1.0	$x_1(x_0 + 1)$	$x_1(x_0 + 1)$	$x_1(x_0 + 1)$
F_{02}	1.0	1.0	1.0	$x_0 + x_1^2 + x_1$	$x_0 + x_1^2 + x_1$	$x_0 + x_1^2 + x_1$
F_{03}	1.0	1.0	1.0	$x_0x_1 + cos(x_0)$	$x_0x_1 + cos(x_0)$	$x_0x_1 + cos(x_0)$
F_{04}	1.0	1.0	1.0	$x_0 - x_1^2$	$x_0 - x_1^2$	$x_0 - x_1^2$
F_{05}	1.0	1.0	1.0	$x_0 - x_1^2 + x_1$	$x_0 - x_1^2 + x_1$	$x_0 - x_1^2 + x_1$

- Sve metode su pronašle izraze koji su ekvivalentni sa ciljnim izrazom.
- Za istu instancu, sve metode su vratile isti izraz.

Evaluacija i poređenje metaheurističkih metoda

Tabela: Informacije o izrazu koji daje maksimalnu R^2 vrednost na test skupu od svih izraza dobijenih pri 30 nezavisnih pokretanja

	Maksimalna R^2 vrednost na test skupu			Simbolička ekvivalencija sa ciljnim izrazom		
	GP	GP sa SSC	VNP	GP	GP sa SSC	VNP
F_1	0.992	0.985	**1.0**	Ne	Ne	Da
F_2	0.994	0.981	**1.0**	Ne	Ne	Da
F_3	0.981	0.995	**1.0**	Ne	Ne	Da
F_4	0.986	0.960	**1.0**	Ne	Ne	Da
F_5	0.943	0.964	**0.999**	Ne	Ne	Ne
F_6	0.971	0.987	**1.0**	Ne	Ne	Da
F_7	0.997	0.998	**0.999**	Ne	Ne	Ne
F_8	0.999	0.994	**1.0**	Ne	Ne	Da
F_9	**1.0**	**1.0**	**1.0**	Da	Da	Da
Yacht	0.929	0.792	**0.956**	/	/	/