UDRUGA HRVATSKIH GRAĐEVINSKIH FAKULTETA
SVEUČILIŠTE U ZAGREBU GRAĐEVINSKI FAKULTET
ZAJEDNIČKI TEMELJI '17.

DOI: https://doi.org/10.5592/CO/ZT.2017.13

Kalibracija modela nadziranog učenja optimizacijom modelskih parametara

Jadran Berbić[1], Eva Ocvirk[2], Dijana Oskoruš[1], Tatjana Vujnović[1]

[1] Državni hidrometeorološki zavod, Hrvatska
[2] Sveučilište u Zagrebu, Građevinski fakultet
kontakt: jberbic@hotmail.com

Sažetak

Pretraživanje optimalnih parametara modela nadziranog učenja u svrhu hidrološkog modeliranja može biti zahtjevno u vremenskom smislu. Trajanje pretraživanja parametara ovisi o reprezentaciji modela, količini primjera za gradnju i kalibraciju modela, količini variranih parametara, rasponu i količini parametara unutar raspona, broju ispitanih kombinacija podjele skupa podataka na dijelove za gradnju i kalibraciju, prirodi modela za odabrane parametre. Ovisnost preciznosti modela o odabiru parametara smanjuje se s povećanjem količine primjera za gradnju modela te je u slučaju smanjene količine primjera potrebno više pažnje posvetiti odabiru parametara. U radu su prikazani rezultati pretraživanja parametara za tri modela nadziranog učenja različitih konfiguracija ulaznih varijabli. Pretraživanje je provedeno pomoću dva osnovna principa: odabirom najboljeg rješenja za odabrane raspone i količinu parametara te korištenjem optimizacijskog algoritma – simuliranog kaljenja.

Ključne riječi: *nadzirano učenje, parametri modela, optimizacijski algoritmi*

Calibration of supervised learning models using model parameters optimization algorithms

Abstract

Searching for optimal parameters of supervised learning models for the purpose of hydrological modelling tend to be timely exhaustive. Duration of parameter searching depend of model representation, number of instances for model training and calibration, number of varied parameters, range and number of parameters in the range, number of used combinations of dataset division in training and calibration part, model nature for chosen parameters. Dependence of model precisions of parameters choice is decreased with increase of number of instances for model training, and in the case of decreased number of instances more attention has to be put on parameter searchin Results of parameter searching for three supervised models with different input variables configuration are shown in the paper. Searching is done by using two basic principles: choice of best solution for chosen ranges and number of parameters and by searching with optimization algorithm – simulated annealin.

Keywords: *supervised learning, model parameters, optimization algorithms*

1 Uvod

Predmetno istraživanje pripada domeni dugoročnih predviđanja u hidrologiji koristeći modele nadziranog učenja u svrhu predviđanja mjesečnog, sezonskog i dugoročnog planiranja. Dugoročno predviđanje nadziranim učenjem predstavlja izazov iz dva osnovna razloga: količina primjera (instanci) za gradnju modela značajno je manja nego u slučaju kratkoročnih predviđanja, izgradnjom odgovarajućih modela doprinosi se dugoročnom planiranju hidrotehničkih sustava. Naime, sa smanjenjem broja primjera smanjuje se vjerojatnost izgradnje kvalitetnog modela jer je nadzirano učenje zasnovano na pronalasku uzoraka u podacima. Planiranje hidrotehničkih sustava podrazumijeva procjenu mogućnosti zadovoljenja o sustavu ovisnih korisnika u dugoročnom smislu (tijekom trajanja sustava). Za takvu procjenu potrebna je informacija o vremenskoj dinamici dostupne količine vode te su modeli nadziranog učenja korišteni za predviđanje srednjeg mjesečnog dotoka. Svaka izgradnja hidroloških modela općenito se sastoji od dva dijela: odabira modelske konfiguracije – ulaznih varijabli te odabira modelskih parametara. Iz pregleda područja zaključeno je da se odabir konfiguracije (ulaznih varijabli) modela nadziranog učenja uglavnom provodi procesom pokušaja i pogreške. Složenost postupka odabira konfiguracije proizlazi iz značajne varijabilnosti u mogućnosti odabira konfiguracije, uz što postoji i značajna varijabilnost parametara modela. Konfiguracije modela korištene u radu dobivene su metodom pokušaja i pogreške uzimajući u obzir koreliranost ulaznih varijabli i predviđane varijable – srednjeg mjesečnog dotoka. Tri konfiguracije modela s ulaznim karakterističnim veličinama redom dotoka, oborine i temperature s jedne meteorološke postaje te oborina i temperature s dvije postaje zapisane su u sljedećem obliku:

$$Q_{avm} = f\ (Q_{avm\text{-}1}, Q_{\text{min-}11}, Q_{\text{min-}1}, Q_{\text{max-}11}, Q_{yavm},\ Q_{av\,\text{min-}11}) \tag{1}$$

$$Q_{avm} = f\ (P_{1,avm\text{-}1}, P_{1,avm}, P_{1,acc\text{-}11}, P_{1,acc\text{-}2}, P_{1,acc\text{-}1}, P_{1,\text{max-}1}, P_{1,avacc\text{-}2}, T_{1,avm\text{-}11}, T_{1,av\,\text{min-}2}) \tag{2}$$

$$\begin{aligned} Q_{\text{avm}} = \text{f}(&P_{1,\text{avm-}1}, P_{1,\text{avm}}, P_{1,\text{acc-}11}, P_{1,\text{acc-}2}, \text{P}_{1,\text{acc-}1}, \text{P}_{1,\text{avacc-}2}, \text{P}_{1,\text{avacc}}, T_{1,\text{avm-}11}, T_{1,\text{avmin-}2}, P_{2,\text{avm-}11}, \\ &P_{2,\text{avm}}, P_{2,\text{acc-}11}, P_{2,\text{acc-}1}, P_{2,\text{acc}}, P_{2,\text{max}}, P_{2,\text{yavm}}, P_{2,\text{avacc-}11}, P_{2,\text{avacc}}, T_{2,\text{avm-}11}, T_{2,\text{avm-}2}, T_{2,\text{avm}}) \end{aligned} \tag{3}$$

Tablica 1. Veličine i oznake korištenih ulaznih varijabli

Veličina	**Q…protok [m³/s]**	**T…temperatura zraka [°C]**	**P…oborina** [mm]
Oznaka	avm, min, max, yavm, avmin, avmax	avm, min, max, yavm, avmin, avmax	avm, acc, max, yavm, avacc, avmax
avm, min, max - srednja, minimalna i maksimalna mjesečna vrijednost; yavm, avmin, avmax - srednja, minimalna i maksimalna mjesečna vrijednost usrednjena po svim godinama; acc - akumulirana mjesečna vrijednost; avacc - akumulirana mjesečna vrijednost usrednjena po svim godinama			

Indeks 1 odnosi se na podatke s glavne meteorološke postaje Knin, a indeks 2 na temperaturu s klimatološke postaje Sinj i oborinu s kišomjerne postaje Vinalić. Dotoci su s hidrološke postaje Vinalić 1. Modeli (reprezentacije) nadziranog učenja ovdje su neuronske mreže (eng. *artificial neural network* - ANN), metoda potpornih vektora (eng. support vector machine - SVM) i metoda najbližih susjeda (eng. *nearest neighbors method* - NNM). NNM, iako ima dosta varijacija u izboru tipa i parametara modela, jednostavniji je model od preostala dva, ali upućuje na to kakva se preciznost može očekivati kod druga dva modela, jer su uz pažljiv odabir parametara druga dva modela obično preciznija. Opis modela i značenja njihovih parametara može se naći u literaturi o strojnom učenju [1-5].

1.1 Metodologija

Modelima nadziranog učenja se na temelju zadanih podataka (ulaznih i izlaznih varijabli) aproksimira funkcija kojom se izvode predviđanja na neviđenim primjerima. Iz same definicije jasno je da na kvalitetu modela utječe duljina korištene vremenske serije (količina instanci) za proces izgradnje, kalibracije i verifikacije modela. Radi procjene utjecaja duljine niza na preciznost modela te dolaska do odgovora na pitanje koje su minimalne duljine niza potrebne za izgradnju modela zadovoljavajuće točnosti, građeni su modeli nadziranog učenja na vremenskim serijama s istim varijablama, ali različite duljine. Korištene duljine niza su od 10 do 65, 67 i 70 godina, ovisno o podacima na raspolaganju kod pojedine konfiguracije modela. Za svaku konfiguraciju modela i korištenu duljinu niza, optimizirani su parametri modela ANN, SVM i NNM. Podaci koji nisu korišteni u procesu izgradnja-kalibracija-verifikacija, odnosno razlika ukupne količine podataka i korištene duljine niza, iskorištena je za proces dodatne verifikacije modela. Primjerice, ako je povijesna duljina niza jednaka 65 godina, a prvih 40 godina korišteno je za proces izgradnja-kalibracija-verifikacija modela (kronološki podijeljeno na 60, 20 i 20 % podataka), ostalih 25 godina iskorišteno je za dodatnu verifikaciju modela. Na taj način ispitana je mogućnost uporabe modela za postupak dugoročnog planiranja kod konfiguracija koje koriste oborinu i temperaturu kao ulazne varijable. U postupku pretraživanja optimalnih parametara korišteni su dijelovi podataka za izgradnju i kalibraciju modela, a maksimizirao se koeficijent determinacije R^2. R^2 je mjera izglednosti predviđanja modelom, načelno u rasponu od 0,0 do 1,0, iako prema nekim definicijama može biti i manji od 0,0 jer model može predviđati proizvoljno loše [3]. Savršeno precizno predviđanje karakterizira vrijednost 1,0 [6]. Kod ANN-a za aktivacijske funkcije tangens hiperbolički (eng. *tanh*) i rektifikacijsku funkciju (eng. *relu*), uz zadani konstantni moment učenja 0,9, variran je broj čvorova u skrivenom sloju *HLS*, početni intenzitet učenja *LRI* i tolerancija greške *TOL*. Kod SVM-a za funkcije kernela – radijalno zasnovanu funkciju (eng. *rbf*), polinom i sigmoidnu funkciju, varirani su parametar razmjene *C*, širina margine ε i parametar kernela γ. Kod NNM-a za način

raspodjele težinskih koeficijenata po susjedima – jednoliki i prema udaljenostima susjeda, variran je broj susjeda *nn*, algoritmi izračuna težinskih koeficijenata (eng. *Brute, Ball tree, Kd tree, Auto*) i potencija Minkowskog za izračun udaljenosti *p*. U poglavlju 2. dani su rezultati pretraživanja provedenog na temelju ugrađenih funkcija u Pythonovoj knjižnici *sklearn*. Za zadanu konfiguraciju modela, raspon i količinu parametara u tom rasponu, rezultat su parametri modela s najvećom vrijednošću R^2. Pri pretraživanju se cijeli dio vremenske serije za gradnju i kalibraciju modela po nasumičnom izboru dijeli na ta dva dijela, 75 % za gradnju, a 25 % za kalibraciju. U konfiguraciji s dotokom korištena je jedna takva podjela u svakom pretraživanju, a u ostale dvije konfiguracije tri. U poglavlju 3. pretraživanje je provedeno programskim rješenjem simuliranog kaljenja. Vremenska serija dijeljena je na dijelove za gradnju i kalibraciju na isti način kao u prethodnom slučaju. Zbog ograničenosti prostora, dan je osvrt samo na modele s duljinom niza od 40 godina, optimalne po preciznosti i veličini dijela za dodatnu verifikaciju kod svih konfiguracija i modela.

2 Detaljno pretraživanje odabranih raspona parametara

U tablici 2. prikazani su rasponi korišteni za pretraživanje za sve konfiguracije. Kod ANN-a pretraživan je broj čvorova s korakom 3, a u svakoj konfiguraciji pretraženo je ukupno 20 različitih vrijednosti *LRI* te 20 vrijednosti *TOL.* Kod SVM-a je za svaki parametar pretraženo po 30 vrijednosti, osim kod konfiguracije 2 gdje je pretraženo 35 vrijednosti parametra ε. Tijekom analize odabir raspona parametara je mijenjan ovisno o tome u kojim rasponima se nalaze najbolji modeli iz prethodno korištene duljine niza.

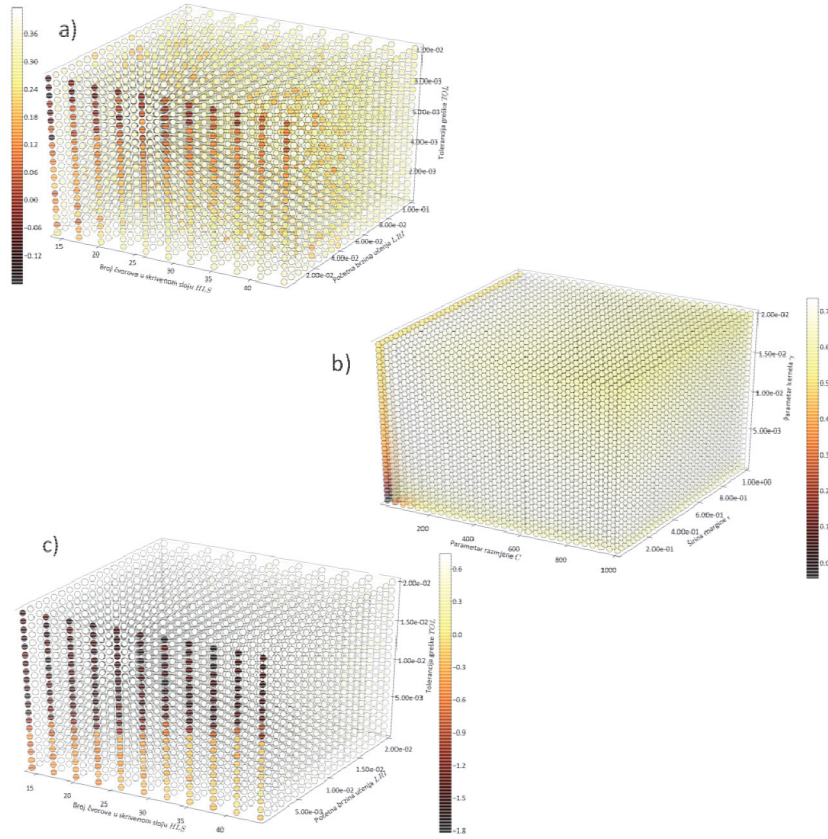
Tablica 2. Korišteni rasponi parametara modela za konfiguracije (1-3)

Konf.	ANN			SVM		NNM		
	HLS	*LRI*	*TOL*	*C*	ε	γ	*nn*	*p*
(1)	14-47	$5{\cdot}10^{-5}$-10^{-1}	$5{\cdot}10^{-5}$-10^{-2}	5,0-750,0	$5{\cdot}10^{-4}$-$8{\cdot}10^{1}$	10^{-3}-$2{\cdot}10^{-2}$	2-30	1-20
(2)	14-47	10^{-5}-10^{-2}	$5{\cdot}10^{-6}$-10^{-2}	1,0-1000,0	10^{-5}-10^{1}	10^{-4}-$2{\cdot}10^{-2}$		
(3)	14-44	10^{-5}-$2{\cdot}10^{-2}$	$5{\cdot}10^{-6}$-10^{-2}	1,0-1000,0	$5{\cdot}10^{-4}$-$8{\cdot}10^{1}$	10^{-5}-$8{\cdot}10^{-3}$		

Postignute vrijednosti R^2 i pripadni parametri prikazani su u tablici 3., a grafovi koji prikazuju preciznost u ovisnosti o parametrima predočeni su na slici 1.

Tablica 3. Optimalni parametri modela za konfiguracije 1-3

Konf.	ANN				SVM				NNM				R^2 [1]		
	Akt. f.	*HLS*	*LRI*	*TOL*	Kernel	*C*	ε	γ	Tež. k.	*nn*	Alg.	*p*	ANN	SVM	NNM
(1)	*relu*	20	0,032	0,0026	*rbf*	698,6	6,34	0,008	udalj.	4	*Ball tree*	8	0,42	0,37	0,37
(2)	*tanh*	14	0,0026	0,0026	*rbf*	35,4	0,559	0,019	udalj.	7	*Ball tree*	3	0,73	0,74	0,64
(3)	*relu*	23	0,0126	0,0116	*rbf*	35,4	0,07	0,0077	jedn.	4	*Ball tree*	2	0,74	0,73	0,69



Slika 1. Preciznost modela za zadane raspone parametara: a) ANN za konfiguraciju 1, b) SVM za konfiguraciju 2, c) ANN za konfiguraciju 3

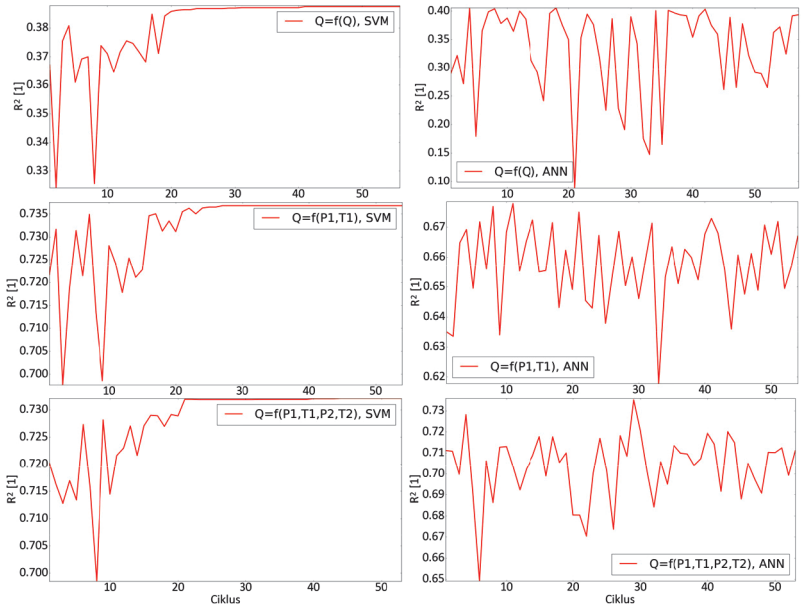
3 Pristup primjenom simuliranog kaljenja

Simulirano kaljenje moderni je optimizacijski algoritam koji primjenjuje analogiju između simulacije termalnog kaljenja i pretraživanja optimalnog rješenja. Princip koji se koristi postupno je hlađenje metala kojim se teži postići minimum unutarnje energije kristalizacijom atoma. Za više informacija čitatelja se upućuje na literaturu [7]. Programsko rješenje napisano je u okruženju Python. Algoritam za zadanu početnu

količinu proizvoljno odabranih parametara u zadanim rasponima pronalazi prosjek vrijednosti ciljne funkcije (R^2) skaliran na odgovarajuće vrijednosti temperature (0 K - 400 K). Za zadani broj iteracija u blizini generiranih slučajnih parametara modela pronalazi nove skupove parametara modela. Ako je ciljna vrijednost manja od one iz prethodne iteracije, ona se prihvaća, a ako je veća, prihvaća se s nasumično generiranom vjerojatnošću. Dakle, problem je po definiciji minimizacijski, ali je kroz parametar temperature preoblikovan u maksimizacijski, tj. smanjenjem temperature R^2 se povećava. Generiranje novih parametara prestaje nakon zadanog broja iteracija, a pretraživanje kreće u novi ciklus te se temperatura smanjuje po unaprijed zadanom pravilu. Pretraga prestaje nakon što se temperatura smanji na zadanu maksimalnu vrijednost ili pri maksimalno dopuštenom broju ciklusa. Pretraživanje parametara provedeno je u istom rasponu kao i detaljno pretraživanje parametara.

Tablica 4. Optimalni parametri modela dobiveni simuliranim kaljenjem

Konfiguracija	ANN				SVM				R^2 [1]	
	Akt.f.	*HLS*	*LRI*	*TOL*	Kernel	*C*	ε	γ	ANN	SVM
(1)	*relu*	28	0,0018	0,0054	*rbf*	653,0	6,134	0,0050	0,40	0,39
(2)	*tanh*	45	0,0006	0,0035	*rbf*	31,5	0,925	0,0166	0,70	0,74
(3)	*relu*	41	0,0043	0,0076	*rbf*	970,2	4,662	0,0028	0,67	0,73



Slika 2. Pretraživanje parametara modela ANN (desno) i SVM (lijevo) korištenjem simuliranog kaljenja za konfiguracije 1-3 (odozgo prema dolje)

4 Zaključak

Metodologija pretraživanja parametara i konfiguracije hidroloških modela nadziranim učenjem nije u potpunosti razjašnjena. Vrlo je širok raspon odabira konfiguracije i parametara. Primjećuje se da je kod dobro odabrane konfiguracije (2 i 3) znatno širi raspon parametara za koje je moguće dobiti modele zadovoljavajuće preciznosti (s koeficijentom determinacije 0,7 - 0,8). Korištenjem optimizacijskog algoritma simuliranog kaljenja kod SVM-a pronađeni su parametri podjednake točnosti, ali uz kraće trajanje pretrage (8,3; 13,9; 10,7 min umjesto 5,4; 47,3; 37,9 min redom za konfiguracije 1; 2; 3). Također, kod SVM-a je algoritam konvergirao, dok se kod ANN-a primjećuje smanjenje oscilacija s povećanjem broja ciklusa, ali nema konvergencije. Time i završni odabir parametara nije optimalan, a preciznost je manja nego kod detaljnog pretraživanja. ANN s većim vrijednostima *LRI* može biti nestabilna (zbog zaobilaženja globalnog minimuma u samoj gradnji modela) te bi trebalo težiti nižim vrijednostima *LRI*, a što opet povećava vrijeme trajanja pretrage jer izgradnja ANN s manjim *LRI* dulje traje. Vrijeme kod pretraživanja simuliranim kaljenjem za ANN trajalo je dulje, redom 330,4; 994,2; 655,8 min umjesto 26,5; 154,6; 461,7. Stoga, automatizacija detaljnog pretraživanja konfiguracije uz pretraživanje optimalnih parametara simuliranim kaljenjem na modelu SVM može uštedjeti vrijeme izgradnje modela nadziranog učenja i pridonijeti izgradnji modela zadovoljavajuće točnosti. Također, može se pokazati da primjena simuliranog kaljenja na SVM-u s vremenskom serijom podijeljenom kronološki na dijelove za gradnju i kalibraciju traje još kraće (oko 1 min - 2 min). Promjenom parametara simuliranog kaljenja te raspona parametara ANN-a može se doći u područje stabilnijih rješenja, ali to neće značajno utjecati na smanjenje trajanja pretraživanja.

Literatura

[1] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression, Statistics and Computing, 14 (2004), pp. 199-222.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Duborg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay M.É.: Scikit-learn: machine learning in Python, Journal of Machine Learning Research, 12 (2011), pp. 2825-2830

[3] Python: Scikit-learn user guide, release 0.17, 2015.

[4] Mitchell, T.M.: Machine learning, 1. izdanje, McGraw Hill Inc., New York, 1997.

[5] Russel, R., Norvig, P.: Artificial Intelligence: A Modern Approach, 3. izdanje, Prentice Hall,

2010.

[6] Marsland, S.: Machine Learning, An Algorithmic Perspective, 2. izdanje, Chapman & Hall, 2015.

[7] Rao, S.S.: Engineering Optimization, 4. Izdanje, John Wiley and Sons, 2009.