Milenko Gavrić

Poređenje predviđanja vodostaja reke na osnovu istorijskih podataka upotrebom neuronske mreže i skrivenog Markovljevog modela

U ovom radu poređeni su rezultati predviđanja vodostaja reke dobijeni pomoću dva različita pristupa veštačkoj inteligenciji, neuronske mreže i Skrivenog Markovljevog modela. Posmatrane su stanice koje mere visinu vodostaja reka Velika Morava i Crnica kod Ćuprije, Paraćina i Varvarina. Korišćeni su podaci vodostaja reka Srbije u periodu od 2004-2012. godine, kao i vremenske prognoze za dati region. Neuronska mreža je trenirana koristeći RPROP (Resilient back PROPagation) algoritam, a predviđanja su poređena koristeci kros-validaciju. Neuronska mreža je koristila visinu vodostaja reke na mernim stanicama Varvarin i Paraćin, kao i vremenske uslove, tj. padavine, temperaturu, vlažnost vazduha i pritisak. Prosečna greška neuronske mreže je 15 cm. Skriveni Markovljev model je koristio samo visinu vodostaja reke za mernu stanicu Ćuprija i njegova prosečna greška je 10 cm.

Uvod

U zadnjih nekoliko godina primena veštačke inteligencije je postala jedna od bitnih novosti IT sektora. Istražene su razne mogućnosti veštačke inteligencije i zaključuje se da je ona primenljiva u mnogim oblastima. Primena veštačke inteligencije može da se odnosi na automatsku vožnju automobile, predviđanju ponašanja berzanskih transakcija, prepoznavanje bolesti, igranje različitih igara (od šaha do igre go), pa sve do automatskog učenja, gde je na primer Googlova AI simulacija naučila samu sebe da hoda po različitim terenima. U literaturi postoje radovi koji su ispitivali mogućnost predviđanja vodostaja reka, na primer rad Marine Campolo i saradnika (Campolo *et al*. 1999), koji je koristio neuronske mreže, ima uspešnost u predviđanju visine vodostaja od 4% do 13%, sa prosekom od 8%, na nivou od jednog sata, dok se greška povećava za duže vremenske intervale, na primer predviđanje na nivou od 5 sati ima prosečnu grešku od 22%. Rad Bunchingiva Bazartserena i saradnika (Bazartseren *et al*. 2003) koji se oslanjao na predviđanje na neuronske mreže i fuzzy neuronske mreže, a kao ulazne parametre koristio merenja svakog sata svakog dana od 1983-1999. godine je imao koren prosečne kvadratne greške od 3.398 na nivou od jednog sata, 4.303 na nivou od 5 sati, 5.821 na nivou od 10 sati i 7.845 na nivou od 15 sati.

Cilj ovog rada je predviđanje vodostaja reke, bez definisanog vremenskog roka, korišćenjem neuronske mreže i Skrivenog Markovljevog modela (SMM), kao i provera primenljivosti SMM za predviđanje visine vodostaja reke. Podaci koju su korišćeni su vodostaj reke, datum, temperatura, pritisak, vlažnost vazduha, padavine, osunčanost. Tokom izrade projekta razvijene su aplikacije u C# programskom jeziku koje koriste Encog biblioteku (http://www.heatonresearch.com/encog/) i u Python programskom jeziku.

Teorijske osnove

U ovom projektu su korišćena dva različita pristupa veštačkoj inteligenciji, neuronske mreže (Artificial neural network) i SMM za predviđanje vodostaja reke.

Milenko Gavrić (1999), Novi Sad, Jirečekova 7, učenik 3. razreda Gimnazije „Jovan Jovanović Zmaj" u Novom Sadu

MENTOR: Dragan Toroman, Istraživačka stanica Petnica

Neuronska mreža

Neuronska mreža predstavlja jedan od načina na koji se implementira veštačka inteligencija. Osnovna ideja neuronske mreže je da prepozna nelinearne funkcije. Njena struktura se, po ugledu na biološki sistem, sastoji od međusobno povezanih neurona. Prilikom implementacije određenog sloja neurona, uvek je data njegova aktivaciona funkcija. Postoje tri vrste neurona i grupisani su u slojeve:

- ulazni (input),
- skriveni (hidden) i
- izlazni (output).

Broj ulaznih i izlaznih slojeva je ograničen, dok je broj skrivenih slojeva neograničen. Međusobne veze neurona imaju definisane svoje početne težine (weights) koje se menjaju tokom obučavanja mreže i koriste se radi prilagođavanja rezultata koje jedan neuron prenosi drugom. Slojeve jednostavne neuronske mreže prikazuje slika 1.

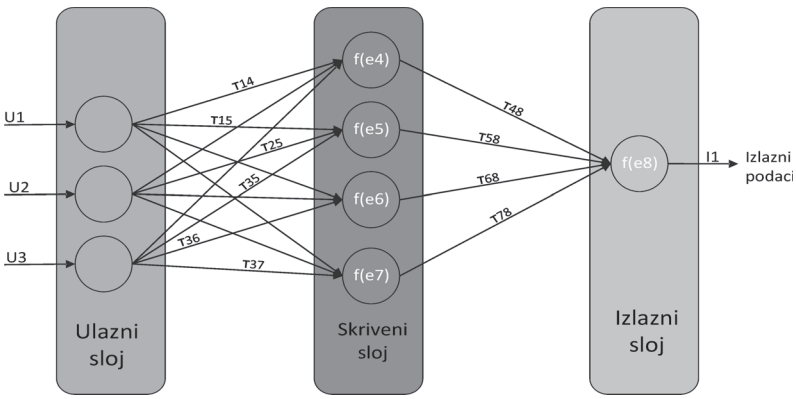
Backpropagation algoritam je izabrani način obučavanja neuronske mreže u ovom radu. Ovaj način obučavanja neuronske mreže radi tako što se za definisane ulazne vrednosti i očekivane rezultate izračunaju izlazni rezultati, koji se kasnije porede sa očekivanim izlaznim rezultatima. Računa se greška između izračunatih i očekivanih rezultata. Na osnovu izračunate greške težina neurona se prilagođava od neurona najbližih izlaznom sloju, pa do neurona najbližih ulaznom sloju.

Skriveni Markovljev model

Skriveni Markovljev model je podvrsta Markovljevih modela koji se koriste za analiziranje stohastičkih sistema. Za svaku vrstu Markovljevih modela važi Markovljevo svojstvo, tj. da trenutna promenljiva zavisi samo od prošle. Markovljev lanac je vrsta Markovljevih modela kod kojih se stanja koja se predviđaju mogu direktno posmatrati. SMM je vrsta Markovljevih modela kod koga Markovljev lanac ne može direktno da se posmatra, ali postoje posmatranja koja na neki nepoznat način utiču na skriveni Markovljev lanac. SMM se sastoji od Markovljevog lanca, u kojem se nalaze skrivene promenljive, i posmatrane promenljive, koje indirektno utiču na skrivena stanja. Ovaj model opisuje se preko tri matrice, emisione, tranzicione i početne.

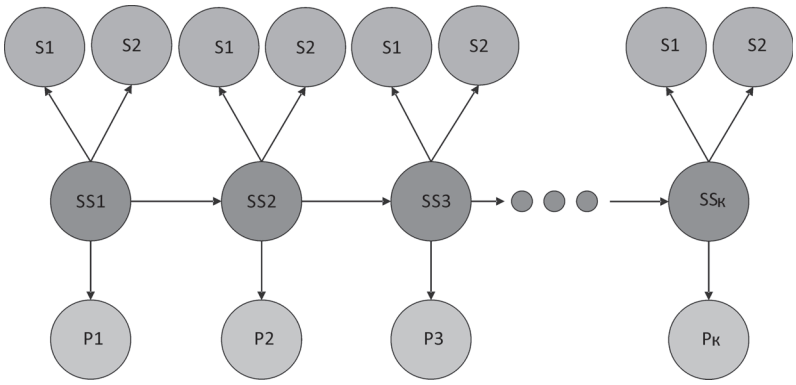
- Tranziciona matrica označava verovatnoće prelaza iz jednog skrivenog stanja u drugo.
- Emisiona matrica označava verovatnoće da određeno posmatrano stanje odgovara određenom skrivenom stanju.
- Početna matrica označava verovatnoće da prvo posmatranje ima stanje h_i.

Primer SMM-a prikazuje slika 2. Oznakom S su označena moguća stanja koje skriveno stanje može da zauzme, oznakom SS su označena skrivena stanja, a oznakom P su označena posmatrana stanja.



Slika 1. Prikaz slojeva jednostavne neuronske mreže

Figure 1. Example of neural network (from left: input, hidden, and output layer)



Slika 2. Prikaz Skrivenog Markovljevog modela

Figure 2. Example of Hidden Markov model

Razlikuju se dve vrste SMM-a: diskretni i kontinualni. Kod diskretnog modela postoji konačan broj posmatranih stanja i emisiona matrica. Primer je vremenska prognoza: vreme može biti kišovito, sunčano ili oblačno. Pošto je ovo SMM, vreme se ne može posmatrati direktno, nego, na primer, preko ljudi. Posmatrane promenljive bi u ovom slučaju predstavljale da li većina ljudi napolju nosi kišobran ili ne. Kod kontinualnog modela ne postoji konačan broj posmatranih stanja, a samim tim ni emisiona matrica. Primer je prepoznavanje govora.

Postoje tri osnovna koraka pri svakoj iteraciji obučavanja Markovljevih lanaca:

1. Ocenjivanje – pronalaženje verovatnoće $p(x|\Theta)$, tj. verovatnoće da dati model generiše posmatrane promenljive X: x_1, x_2, …, x_n. Ovo se postiže forward-backward algoritmom.
2. Dekodiranje – pronalaženje najverovatnije sekvence skrivenih stanja modela Θ čiji su proizvodi posmatrane promenljive X : x_1, x_2, …, x_n. Ovaj problem se rešava Viterbi algoritmom.
3. Učenje – prilagođavanje parametara modela $\{T, E, \pi\}$ radi maksimizovanja verovatnoće $p(x|\Theta)$ za dati model Θ^{old} i sekvencu posmatranih promenljivih X: x_1, x_2, …, x_n tako da novodobijen model Θ^{new} daje najbolje moguće rezultate. Ovo se postiže na dva načina, Maximum likelihood i Maximum mutual information kriterijumom.

Kontinualni skriveni modeli imaju beskonačan broj mogućih skrivenih stanja. Pošto je veličina emisione matrice definisana preko broja skrivenih stanja i mogućih posmatranih promenljivih, nemoguće je predstaviti takvu matricu u računaru. Zato se ta stanja najčešće zamenjuju nekim raspodelama ili se kvantifikuju. U slučaju da se zamenjuju nekom raspodelom, emisiona matrica se zamenjuje odgovarajućim matricama koje opisuju te raspodele.

Expectation Maximization (EM) algoritam služi za treniranje Markovljevog lanca i SMM-a. Cilj je da se pronađu što bolji parametri modela, tako da se dobije najveća moguća verovatnoća da se predvidi data sekvenca posmatranih promenljivih. Algoritam se sastoji od više koraka. Prvi je Forward-Backward korak u kojem se izračunava verovatnoća da su sva prethodna stanja imala određenu vrednost (forward) i da sva buduća stanja takođe imaju određenu vrednost (backward). Nakon toga se prilagođavaju početna, tranziciona i emisiona matrica, tim redom. Algoritam koji se koristi za izračunavanje Maximum Likelihood kriterijuma je Baum-Welch algoritam.

Opis implementacije

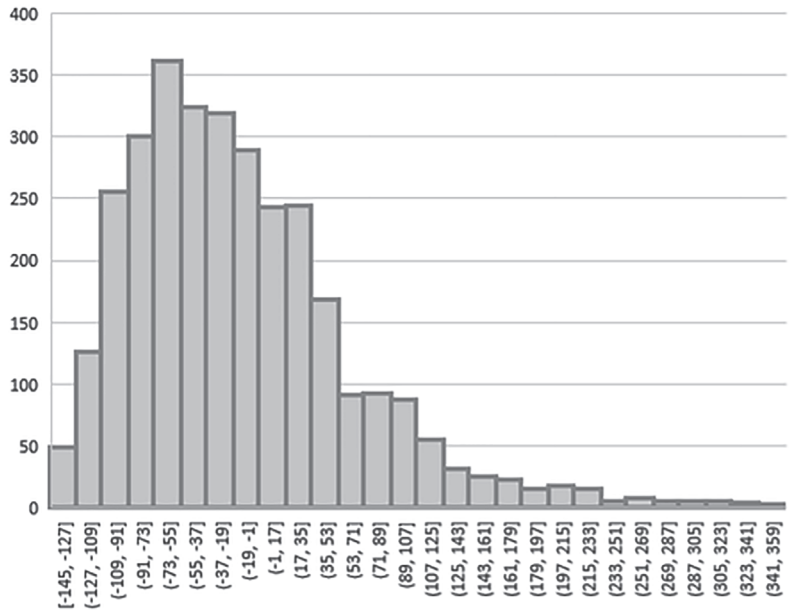
Aplikacija je razvijena u programskom jeziku C#. Ulazni podaci su učitavani iz .txt fajla, nakon čega su podeljeni u 5 grupa jednakih veličina. Ulazni podatak se sastojao od 8 različitih vrednosti. Te vrednosti su visina vodostaja Velike Morave i Crnice, temperature, vlažnost vazduha, pritisak, padavine. Korišćene su dve biblioteke: Encog (http://www.heatonresearch.com/encog/) i Accord.Statistics (http://accord-framework.net/). Encog biblioteka je korišćena za obučavanje neuronske mreže, dok je Accord.Statistics korišćen za SMM. Kreirana je instanca klase BasicNetwork. U datoj instanci su

kreirani skriveni sloj sa 8 neurona i izlazni sloj sa jednim neuronom, sa linearnom funkcijom kao njihovom aktivacionom funkcijom. Zatim je neuronska mreža trenirana na osnovu ulaznih podataka koristeći ResilientPropagation algoritam.

SMM je prvobitno implementiran uz oslonac na biblioteku Accord.- Statistics. Aplikacija je napisana u programskom jeziku C#, a korišćene su klase HiddenMarkovModel i BaumWelch Learning. Kao ulaz su korišćeni podaci visine vodostaja reke iz stanice Ćuprija. Korišćena je metoda Predict klase HiddenMarkovModel. Implementirana metoda nije davala očekivane rezultate, tj. predviđala je samo u slučaju da sekvencu poznaje. U slučaju da ima više sličnih sekvenci, birala je onaj element koji se najviše puta ponavljao, inače je vraćala kao rezultat 0. Ovo nije odgovaralo očekivanom izlazu te funkcije, koja je trebalo da „nauči" pravilo u datoj sekvenci i primeni ga. Sa obzirom da metoda nije davala očekivane rezultate za nepostojeće sekvence, ona nije mogla biti korišćena u ovom radu, te se odustalo od daljeg rada sa SMM modelom. Nakon toga je model implementiran u programskom jeziku Python. Pošto SMM analizira sekvencu, kao ulaz su mu date visine vodostaja reke na mernoj stanici Ćuprija, i to u periodu od mesec dana. Ako je zadat duži vremenski period za analiziranje, model nije radio zbog aproksimacije verovatnoće da se određena sekvenca posmatra. Postavljena su dva skrivena stanja, od kojih je svako skriveno stanje opisano Gausovom mešovitom raspodelom. Svaka Gausova mešovita raspodela je u sebi sadržala dve Gausove raspodele. Centri tih Gausovih raspodela su dobijeni tako što je izračunat prosek visine vodostaja reke koje se analizira, i on je smanjen, tj. povećan za određenu vrednost. Kao rezultat predviđanja se dobija Gausov mešoviti model koji opisuje verovatnoću za sledeću posmatranu promenljivu. Napisana je metoda za semplovanje koja uzima visinu vodostaja reke od prošlog dana i približava je za određeni koeficijent ka centru najbliže Gausove raspodele.

Dobijeni rezultati

Kao testni podaci korišćena su merenja vodostaja od početka 2004. godine do kraja 2012. godine, izvršena od strane RHMZ, kao i vremenske prognoze za region i dati period (istorijski podaci visine vodostaja za reke Srbije od 2004. do 2012. godine dati su na Web 1, a vremenske prognoze za odgovarajući region nalaze se na https://www.ogimet. com/). Testni podaci su



Slika 3. Promena vodostaja kod stanice Ćuprija od 2004. do 2012. godine

Figure 3. Measured water level at Ćuprija station from 2004 until 2012

grupisani tako da su 2 godine služile za proveru, a preostalih 7 godina za trening modela. Na ovaj način je definisano 8 međusobno različitih testnih grupa. Sledeća slika prikazuje broj ponavljanja izmerenih visina vodostaja reke tokom 8 godina na mernoj stanici Ćuprija.

Problemi koji su mogući uz nedovoljno, odnosno preveliko treniranje mreže, su underfitting odnosno overfitting, tj. slučaj kada neuronska mreža nauči „napamet" sve primere. Da bi se ovo izbeglo, korišćen je metod krosvalidacije, tj. test podaci su podeljeni u grupe, nakon čega je neuronska mreža trenirana na svim grupama, osim na jednoj, koja je korišćena kao test.

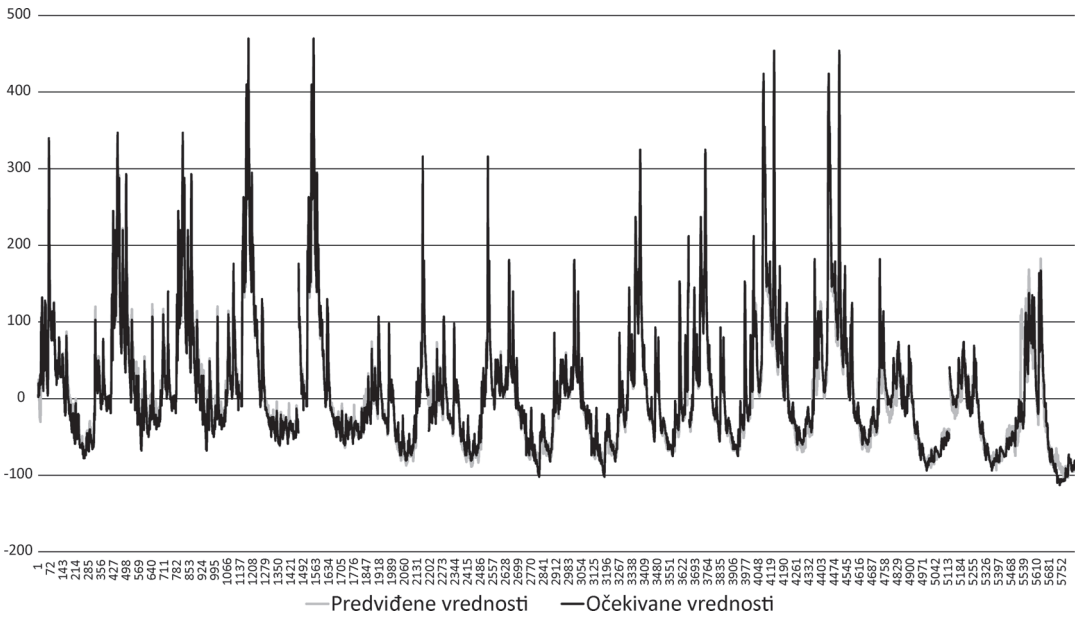
Dobijeni rezultati za greške neuronske mreže pikazani su u tabeli 1.

Razliku između predviđanja neuronske mreže i očekivanih vrednosti visine vodostaja Velike Morave u 2004. godini na stanici Ćuprija prikazuje grafik na slici 4.

Tabela 1. Greške neuronske mreže po test grupama u cm

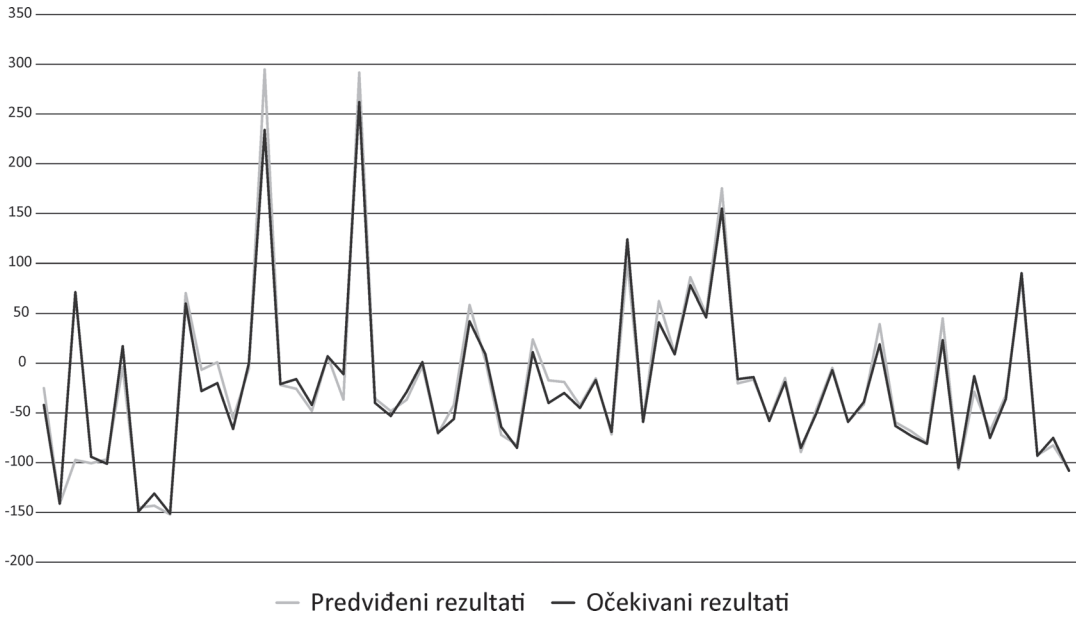
Test grupa	Prosečna kvadratna greška [cm]
Test grupa 1 (2004-2005 godina)	15.38
Test grupa 2 (2005-2006 godina)	14.33
Test grupa 3 (2006-2007 godina)	11.12
Test grupa 4 (2007-2008 godina)	10.03
Test grupa 5 (2008-2009 godina)	11.26
Test grupa 6 (2009-2010 godina)	17.49
Test grupa 7 (2010-2011 godina)	16.40
Test grupa 8 (2011-2012 godina)	23.57
Prosek	15.03

Prosečna greška SMM-a je 10 cm. Kao rezultat, SMM daje raspodelu koja opisuje verovatnoće da se visina vodostaja reke posmatra sutra. Potrebno je semplovati vrednost iz ove raspodele. Za razliku od neuronske mreže, SMM



Slika 4. Predviđanja neuronske mreže u odnosu na očekivane vrednosti za mernu stanicu Ćuprija u periodu od 2004. do 2012. godine u cm

Figure 4. Neural network forecasted values (gray) versus expected values (black) for Ćuprija measurement station from 2004 to 2012 year in cm



Slika 5. Predviđanja Skrivenog Markovljevog modela u odnosu na očekivane vrednosti za mernu stanicu Ćuprija u periodu od 2004. do 2012. godine u cm

Figure 5. Hidden Markov model forecasted values (gray) versus expected values (black) for Ćuprija measurement station from 2004 to 2012 year in cm

radi sa mnogo manje podataka. Koristi samo visinu vodostaja reke na traženoj stanici i mogu se praviti dalja predviđanja na osnovu predviđenih vrednosti jer SMM ima samo jedan ulazni parametar. Grafik na slici 5 prikazuje odstupanja predviđene vrednosti od očekivane i to za svaki mesec. Neki meseci su izbačeni zbog nepostojanja odgovarajućih merenja.

Diskusija i budući rad

Rezultati ovog rada su dve aplikacije, jedna implementirana u C# programskom jeziku koja implementira neuronsku mrežu, a druga u programskom jeziku Python koja implementira SMM. Neuronska mreža je obučavana na podacima između kojih je vremenski razmak jedan dan. Pošto je razmak između dva merenja jedan dan, algoritam se nije mogao testirati na kraćim vremenskim intervalima.

Predviđeni rezultati su upoređeni sa stvarnim merenjima i izračunata je greška. Predviđanja na osnovu predviđenih vodostaja nisu rađena, jer bi to zahtevalo i implementaciju predviđanja ostalih parametara, tj. padavina, temperature, pritiska i vlažnosti vazduha.

Jedan od problema SMM-a jeste što veoma loše predviđa ekstreme. Pošto visinu vodostaja reke opisuje Gausova raspodela, verovatnoća za ekstremne vrednosti je veoma mala, pa samim tim će se veoma retko i predviđati. Takođe postoji problem semplovanja iz dobijene raspodele. Ovo znači da je neuronska mreža bolja za predviđanje ekstremnih vrednosti, dok je SMM bolji za predviđanje vrednosti bližih prosečnim.

U budućim radovima bi se mogle implementirati aplikacije koje mogu da prave naredna predviđanja na osnovu svojih predviđenih vrednosti i aplikacije koje mogu bolje semplovati iz dobijene distribucije.

Literatura

Bazartseren B., Hildebrandt G., Holz K. P. 2003. Short-term water level prediction using neural networks and neuro-fuzzy approach. *Neurocomputing*, **55** (3): 439.

Campolo M., Andreussi P., Soldati A. 1999. River flood forecasting with neural network model. *Water Resource Research*, **35** (4): 1191.

Mitra P., Ray R., Chatterjee R., Basu B., Saha P., Raha S., Barman R., Patra S. 2016. Flood forecasting using Internet of things and Artificial Neural Networks. U *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, Canada, 13–15 October 2016*. IEEE, str. 1–5.

http://accord-framework.net/. Accord biblioteka.

http://www.heatonresearch.com/encog/. Encog biblioteka

https://www.ogimet.com/. Vremenske prognoze.

Web 1. https://drive.google.com/drive/folders/0B8LIzObaDbe-MXpMbWFmaS00d3M

Milenko Gavrić

Comparison of Hidden Markov Model and Neural Network Application for Water Level Forecasting

The goal of this paper is to test the applicability of a continuous hidden Markov model and a suggested method for flood forecasting and its comparison to a neural network. The water levels of the rivers Great Morava and Crnica, from three stations (Varvarin, Paraćin and Ćuprija), from 9 years (2004 to 2012), were analyzed. Two different approaches to artificial intelligence were implemented: neural network and hidden Markov model. In its forecasting, the neural network used the height of the water level measured at Varvarin and Paraćin, as well as the weather forecast for the given region, i.e. precipitation, temperature, humidity and pressure. During the training of the neural network, a method called cross-validation was used. The hidden Markov model used only the height of the water level at the measuring station Ćuprija. After the forecasted values, the two approaches were compared to the expected value and the root mean square error, which is less tolerant to higher errors, was calculated. The root mean square error for the neural network is 15 cm, whereas the root mean square error for the hidden Markov model is 10 cm. Even though the hidden Markov model has a smaller error, the conclusion is that it cannot be used for flood forecasting. To the best of the author's knowledge, currently there is no method in existence that uses data to predict from the given Gaussian density function. Thus, the conclusion is that a neural network is better for forecasting floods.