Elektrotehniški vestnik 76(4): 240–245, 2009
Electrotechnical Review, Ljubljana, Slovenija

Modeliranje časovnih vrst z metodami teorije informacij

Marko Bratina[1], **Andrej Dobnikar**[2], **Uroš Lotrič**[2]

[1] *Savatech, d.o.o., Škofjeloška 6, Kranj, Slovenija*
[2] *Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, Ljubljana, Slovenija*
E-pošta: marko.bratina@sava.si, andrej.dobnikar@fri.uni-lj.si, uros.lotric@fri.uni-lj.si

Povzetek. V prispevku so analizirane možnosti uporabe mer, ki izhajajo iz teorije informacij pri iskanju značilk in modeliranju dinamičnih sistemov iz časovnih vrst. Pri iskanju značilk smo uporabili informacijsko-teoretično zasnovano analizo neodvisnih osi in metodo najznačilnejših projekcij. Pri gradnji modela smo v nevronski mreži, dvoplastnem perceptronu, kot kriterijsko funkcijo poleg najpogosteje uporabljane minimizacije povprečne kvadratne napake uporabili še maksimizacijo informacijskega potenciala. Rezultati napovedovanja na več časovnih vrstah so pokazali, da se predlagane metode iskanja značilk v večini primerov obnašajo bolje kot klasične metode, medtem ko informacijski potencial kot kriterijska funkcija upočasni konvergenco.

Ključne besede: modeliranje časovnih vrst, predprocesiranje, nevronske mreže, teorija informacij

Time-series modeling using information-theory techniques

Extended abstract. The paper analyzes the possibility of using measures originating from the information theory in modeling dynamic systems from time series. Our modeling was based on a multilayered perceptron neural network into which several information measures were integrated.

The presented information measures are based on the Shannon's definitions of entropy and the divergence given in Eqs. (1) and (2). To simplify the extremely computationally intensive methods, the Renyi's counterparts defined by Eqs. (6) and (7) are preferred instead. For example, the information potential given in (9) can be calculated even without integrations. Besides the information potential, the Renyi's approximations of the measures presented by Eqs. (3) and (5) are very useful in modeling. The measures originating from the information theory were included in two ways: in the data preprocessing and as the criterion function during the process of neural network learning.

The basic aim of preprocessing is to reduce the number of inputs to the model and/or reformulate them and thus improve the model performance and generalization ability. Besides classical methods for feature extraction, including the appropriate number of the last values in a time series that results in the best model (H1), the greedy algorithm for finding the combination of inputs, enabling the best model performance (H2) or making a linear combination of original inputs using the principal component analysis (PCA), two methods based on the information theory were tested: the independent component analysis (ICA) and the maximally discriminative projection (MDP).

The neural network learning is usually based on minimizing the mean square error $\mathcal{E}(\mathbf{E})$. As an alternative, maximizing information potential $V_R(\mathbf{E})$ is considered as the learning criterion.

The proposed ideas were tested on five different time series. Fig. 1 gives results for predicting the future value in the time series, while average results are summarized in Table 1. The model performance is estimated by several measures: the normalized root mean square error (NRMSE), the normalized information potential (NIP) and the number of model parameters ($\mathrm{N_{PAR}}$). The ICA preprocessing combined with learning based on information potential yields results comparable to method H2. The latter is, however, computationally extremely intensive. In Fig. 2 and Table 2, classification of the predicted values of time series is presented. As a performance measure, the proportion of correct classification ($\mathrm{P_{OK}}$) is added. The ICA and MDP methods perform equally well in this case.

We show that the methods based on the information theory can be efficiently used in retrieval of relevant features from data. Besides, the usage of the information potential as a criterion function in the neural-network learning process is also promising.

Key words: time-series modeling, pre-processing, neural networks, information theory

1 Uvod

Splošni modeli dinamičnih sistemov poskušajo izluščiti pomembne lastnosti procesov neposredno iz obnašanja merljivih parametrov skozi čas. Temeljijo na predpostavki, da bodo značilnosti, opažene v preteklosti, obstajale še naprej. Pred samim modeliranjem vhodne podatke pogosto pripravimo, na primer s filtriranjem ali metodami za iskanje značilk, s čimer poskušamo poenostaviti modeliranje. Metode predprocesiranja so zelo različne – od popolnoma statističnih do takšnih, ki temeljijo na uspešnosti končnega modela [6].

Večina splošno uveljavljenih metod za analizo in

Prejet 5. maj, 2009
Odobren 19. junij, 2009

modeliranje sistemov v veliki meri predvideva statistiko drugega reda. V zadnjem času se za reševanje čedalje zahtevnejših problemov uveljavljajo rešitve, ki presegajo njene omejitve [1]. Zmogljivosti sodobnih računalniških sistemov so omogočile uporabo idej s področja informacijske teorije tudi pri modeliranju dinamičnih sistemov [2]. Osnove teorije informacij izhajajo s konca prve polovice 20. stoletja, ko je Shannon postavil matematično teorijo za obravnavanje temeljnih vidikov komunikacijskih sistemov. Definicije osnovnih mer, to je entropije, divergence in povprečne medsebojne informacije, izhajajo iz verjetnostne teorije in statistike. Vsaka mera v svojem kontekstu opredeljujejo količino informacije v naključnih spremenljivkah, zato jih lahko učinkovito uporabimo tudi na drugih področjih. Na primer, entropijo in divergenco kot cenilno funkcijo pri modeliranju s splošnimi modeli [3], povprečno medsebojno informacijo pa kot mero za medsebojno povezanost podatkov [4, 5].

V delu se bomo osredinili na problem izbiranja vplivnih podatkov za potrebe modeliranja in na problem samega modeliranja nelinearnih dinamičnih sistemov z metodami, zasnovanimi na teoriji informacij. V drugem poglavju bomo predstavili osnovne koncepte in razširitve informacijske teorije, v tretjem poglavju pa bomo nakazali, kako jih je mogoče uporabiti pri predprocesiranju podatkov. V četrtem poglavju bomo na kratko predstavili, kako lahko v dvoplastnem perceptronu kot kriterijsko funkcijo uporabimo mero, ki izhaja iz teorije informacije. Predstavljeni koncepti bodo nato v petem poglavju tudi praktično ovrednoteni na problemih napovedovanja znanih časovnih vrst. Nazadnje bodo v zaključku povzete prednosti in slabosti uporabe konceptov informacijske teorije v primerjavi z že uveljavljenimi.

2 Mere informacijske teorije

Osnovni meri, ki izhajata iz teorije informacij, sta entropija in divergenca [2]. Shannonova entropija, ki meri nedoločenost naključnega vektorja $\mathbf{X}$ z gostoto verjetnostne porazdelitve $p(\mathbf{x})$, je definirana kot

$$H(\mathbf{X}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \quad . \tag{1}$$

Podobno Kullback-Lieblerjeva divergenca meri podobnost med pravo gostoto porazdelitve $p(\mathbf{x})$ naključne spremenljivke $\mathbf{X}$ in njeno oceno $r(\mathbf{x})$,

$$D(p;r) = -\int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{r(\mathbf{x})} d\mathbf{x} \quad . \tag{2}$$

Medsebojno povezanost spremenljivk X^i z gostotami verjetnostnih porazdelitev $p_i(x^i)$, ki sestavljajo vektor $\mathbf{X} = (X^1, \ldots, X^N)^{\mathrm{T}}$, lahko ocenimo z mero [7]

$$\begin{aligned} J(\mathbf{X}) &= D\left(p(\mathbf{x}); \prod_{i=1}^{N} p_i(x^i)\right) \\ &= \sum_{i=1}^{N} H(X^i) - H(\mathbf{X}) \quad . \end{aligned} \tag{3}$$

Kadar nas zanima količina informacije, ki jo o naključni spremenljivki Y z verjetnostno porazdelitvijo $q(y)$ vsebuje naključni vektor $\mathbf{X}$, je ta mera kar enaka medsebojni informaciji

$$\begin{aligned} I(\mathbf{X};Y) &= H(\mathbf{X}) + H(Y) - H(\mathbf{X},Y) && (4) \\ &= H(\mathbf{X}) - H(\mathbf{X}|Y) \quad , && (5) \end{aligned}$$

pri čemer je $H(\mathbf{X},Y)$ skupna nedoločenost vektorja $\mathbf{X}$ in spremenljivke Y, $H(\mathbf{X}|Y)$ pa povprečna nedoločenost vektorja $\mathbf{X}$, če poznamo spremenljivko Y.

Računanje zgornjih mer je zaradi integralov, ki nastopajo v definicijah, izredno zahtevno. Zato se namesto Shannonove entropije pogosto uporablja Renyijeva entropija [7],

$$H_R(\mathbf{X}) = \frac{1}{1-\alpha} \log \int p^{\alpha}(\mathbf{x}) d\mathbf{x} \quad , \tag{6}$$

ki ji je v limiti $\alpha \to 1$ enaka. Analogno v limiti $\alpha \to 1$ Renyijeva divergenca

$$D_R(p;r) = \frac{1}{\alpha - 1} \log \int p(\mathbf{x}) \left(\frac{p(\mathbf{x})}{r(\mathbf{x})}\right)^{\alpha - 1} d\mathbf{x} \tag{7}$$

preide v Shannonovo. V nadaljevanju se bomo omejili na obravnavanje kvadratnih Renyijevih mer s parametrom $\alpha = 2$. V nasprotju s Shannonovimi merami nobena od relacij (4) in (5) ne velja za Renyijeve mere [5], zato so v nadaljevanju mere, ki izhajajo iz omenjenih relacij, označene z apostrofom.

V praksi se gostota verjetnostne porazdelitve zvezne spremenljivke največkrat oceni z metodo Parzenovega okna [8]

$$p(\mathbf{X}) = K^{-1} \sum_{k=1}^{K} G\left(\mathbf{x} - \mathbf{x}_k\right) \quad , \tag{8}$$

kjer so $\mathbf{x}_i$, $i = 1, \ldots, K$ naključne vrednosti (meritve) vektorja $\mathbf{X}$ in $G(\mathbf{x})$ izbrana jedrna funkcija. Pogosto za jedrno funkcijo uporabimo Gaussovo funkcijo $G(\mathbf{x}) = \prod_{i=1}^{N} G_{h_i}(x^i)$, kjer je $G_{h_i}(x^i)$ enodimenzionalna Gaussova porazdelitev. Ena od možnosti je, da širino porazdelitve določimo s Silvermanovo oceno [9] $h_i = 1{,}06 \sigma_i K^{-0,2}$, kjer je σ_i standardno odstopanje spremenljivke X^i od povprečja.

Ko v definicijo Renyijeve entropije (6) vstavimo nastavek Parzenovega okna (8) ter upoštevamo, da je integral produkta dveh Gaussovih funkcij Gaussova funkcija razlike srednjih vrednosti osnovnih Gaussovih funkcij z dvojno varianco, vidimo, da Renyijev informacijski potencial

$$V_R(\mathbf{X}) = \int p^2(\mathbf{x}) d\mathbf{x} = \int \left(\frac{1}{K} \sum_i G_h(\mathbf{x} - \mathbf{x}_i)\right)^2 d\mathbf{x}$$

$$= \frac{1}{K^2} \sum_j \sum_i G_{h\sqrt{2}}(\mathbf{x}_j - \mathbf{x}_i) \quad (9)$$

in s tem entropijo $H_R(\mathbf{X}) = -\log V_R(\mathbf{X})$ izračunamo brez računsko požrešnih integracij [7].

3 Iskanje značilk

Z metodami za iskanje značilk poskušamo iz vhodnih podatkov izluščiti najpomembnejše značilnosti ter s tem poenostaviti modeliranje in izboljšati odzivnost modelov in njihovo sposobnost posploševanja. Metode se med seboj ločijo po kriterijski funkciji, s katero izbiramo značilke – ta je lahko zasnovana na lastnostih značilk ali pa kar na uspešnosti modela.

3.1 Klasične metode

Značilne spremenljivke lahko izbiramo s pomočjo spektralne in kovariančne analize, ali pa z bolj ali manj intenzivnim preiskovanjem mogočih naborov z različnimi hevrističnimi metodami ali iskalnimi postopki, na primer evolucijskimi [10]. V nadaljevanju se bomo omejili na dve hevristični metodi. Pri prvi (H1) bomo za značilke izbrali določeno število zadnjih spremenljivk v časovni vrsti. Število bomo določili tako, da bo modeliranje čim uspešnejše. Pri drugi metodi (H2) bomo nabor značilk gradili postopoma. V vsakem koraku bomo med značilke vključili tisto od preostalih spremenljivk, ki bo skupaj z že izbranimi značilkami pripeljala do najboljšega modela.

Značilke lahko sestavimo tudi kot linearne kombinacije vhodnih spremenljivk. To omogoča metoda glavnih osi (*ang.* Principal Component Analysis, PCA). Gre za matematični postopek [11], v katerem se iz osnovnih vhodnih spremenljivk sestavi manjše število neodvisnih značilk ali glavnih osi. Prva glavna os je postavljena tako, da ima največjo mogočo varianco. Podobno je vsaka naslednja glavna os postavljena tako, da ima največjo varianco na preostalih podatkih. Ponavadi pri nadaljnjem modeliranju uporabimo nekaj prvih glavnih osi. V nadaljevanju smo izbrali glavne osi, pri katerih je varianca večja od 1 % variance glavne osi.

3.2 Analiza neodvisnih komponent

Analiza neodvisnih komponent (*ang.* Independent Component Analysis, ICA) [12] predvideva, da lahko izmerjene signale $\mathbf{z}_k$ zapišemo kot linearno mešanico statistično neodvisnih signalov, $\mathbf{s}_k$, $\mathbf{z}_k = \mathbf{A}\mathbf{s}_k$. Z analizo neodvisnih komponent želimo poiskati tako matriko $\mathbf{B}$, da bodo značilke $\mathbf{x}_k = \mathbf{B}\mathbf{z}_k = \mathbf{B}\mathbf{A}\mathbf{s}_k$ kar najboljši približek statistično neodvisnih signalov $\mathbf{s}_k$. V prvem koraku s transformacijo $\mathbf{z}'_k = \mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}\mathbf{z}_k$, v kateri je $\mathbf{D}$ diagonalna matrika lastnih vrednosti, $\mathbf{V}$ pa matrika lastnih vektorjev avtokorelacijske matrike vektorjev $\mathbf{z}_k$, poskrbimo, da je avtokorelacijska matrika vektorjev $\mathbf{z}'_k$ identiteta. V drugem koraku nato poiščemo rotacijsko matriko $\mathbf{R}$, ki nam da tak vektor signalov $\mathbf{x}_k = \mathbf{R}\mathbf{z}'_k$, pri katerem je izbrana kriterijska funkcija optimalna. Če želimo, da imajo vektorji $\mathbf{x}_k$ manjše število komponent kot vektorji $\mathbf{z}_k$, na slednjih pred opisanim postopkom uporabimo metodo glavnih osi.

Obstaja več metod analize neodvisnih komponent [13, 14], nekaj jih temelji na teoriji informacije. Če vektorje $\mathbf{x}_k$ razumemo kot točke naključnega vektorja $\mathbf{X}$, potem lahko za kriterijsko funkcijo uporabimo Renyijevo medsebojno informacijo vektorja, ki izhaja iz (7). Ker pa je njen izračun časovno zahteven, se ponavadi uporabi kar zveza (3), v kateri namesto Shannonove uporabimo Renyijevo entropijo [5]. Skupna entropija vektorja $\mathbf{X}$ je invariantna na rotacije [15] , zato lahko zadnji člen v (3) izpustimo in minimiziramo samo vsoto entropij posamičnih signalov $J'_R(\mathbf{X}) = \sum_{i=1}^{N} H_R(X^i)$. Rotacijsko matriko, ki minimizira mero $J'_R(\mathbf{X})$, iščemo iterativno, na primer po postopku najhitrejšega sestopa, po katerem je sprememba rotacijske matrike enaka $\Delta\mathbf{R} = -\eta\, \partial J'_R(\mathbf{X})/\partial\mathbf{R}$ [5].

3.3 Metoda najznačilnejših projekcij

Tako kot pri analizi neodvisnih komponent tudi pri metodi najznačilnejših projekcij (*ang.* Maximally Discriminative Projections, MDP) iščemo značilke $\mathbf{x}_k$, ki so linearne kombinacije osnovnih meritev $\mathbf{z}_k$ z dvofaznim postopkom [16]. Bistvena razlika med metodama je v zasnovi kriterijske funkcije. Medtem ko je pri analizi neodvisnih komponent ta odvisna samo od vhodnih podatkov, jo pri metodi najznačilnejših projekcij določajo odvisnosti med vhodnimi podatki in ustreznimi izhodi iz modela. Opisanemu konceptu ustreza maksimizacija mere medsebojne informacije med vhodnimi in izhodnimi spremenljivkami.

Računanje prave Renyijeve medsebojne informacije je preveč zahtevno, zato se uporablja približek, ki izhaja iz (5), v kateri Shannonove entropije zamenjamo z Renyijevimi. Ta mera se še dodatno poenostavi v primeru, ko želimo z modelom vhodne podatke uvrščati v C vnaprej podanih razredov. Takrat lahko zapišemo

$$\begin{aligned} I'_R(\mathbf{X};Y) &= H_R(\mathbf{X}) - H_R(\mathbf{X}|Y) \\ &= H_R(\mathbf{X}) - \sum_{c=1}^{C} \frac{n_c}{n} H_R(\mathbf{X}|Y{=}\,c)\,, \end{aligned} \quad (10)$$

kjer je n_c število podatkov, ki se uvrščajo v razred c, $H_R(\mathbf{X}|Y=c)$ pa nedoločenost vektorja $\mathbf{X}$ pri uvrščanju v ta razred. Pri računanju entropije $H_R(\mathbf{X})$ uporabimo vse vzorce, pri računanju entropije $H_R(\mathbf{X}|Y=c)$ pa le vzorce, ki spadajo v razred c. Tako kot pri analizi neodvisnih komponent rotacijsko matriko popravljamo iterativno glede na vrednosti gradienta kriterijske funkcije [16].

4 Modeliranje z nevronskimi mrežami

Nevronske mreže spadajo med splošne nelinearne modele, ki izbrano vrednost v časovni vrsti povezujejo s predhodnimi vrednostmi. Nevronska mreža dvoplastni perceptron [17], ki smo jo uporabili pri analizi, ima v skriti plasti M_H nelinearnih nevronov, v izhodni plasti pa M_O linearnih nevronov. Odziv modela lahko opišemo z enačbama

$$\mathbf{x}_k^H = \tanh\left(\mathbf{W}_H \mathbf{x}_k + \mathbf{b}_H\right) \quad \text{in} \tag{11}$$

$$\mathbf{x}_k^O = \mathbf{W}_O \mathbf{x}_k^H + \mathbf{b}_O \quad . \tag{12}$$

Z učenjem na parih vhodno-izhodnih vzorcev $(\mathbf{x}_k, \mathbf{d}_k)$, $k = 1, \ldots, K$ proste parametre modela, uteži $\mathbf{W}_H$ in $\mathbf{W}_O$ ter pragove $\mathbf{b}_H$ in $\mathbf{b}_O$ nastavimo tako, da optimiziramo izbrano kriterijsko funkcijo na podlagi napak $\mathbf{e}_k = \mathbf{d}_k - \mathbf{x}_k^O$, $k = 1, \ldots, K$ med dejanskimi in izračunanimi vrednostmi. Ponavadi minimiziramo povprečno kvadratno napako $\mathcal{E}(\mathbf{E}) = \frac{1}{N_O K} \sum_{k=1}^{K} \mathbf{e}_k^{\mathrm{T}} \mathbf{e}_k$. Med merami, ki izhajajo iz informacijske teorije, je primerna minimizacija nedoločenosti napake. Ob predpostavki, da napake $\mathbf{e}_k$ tvorijo naključni vektor $\mathbf{E}$, se za kriterijsko funkcijo največkrat uporablja minimizacija entropije oziroma maksimizacija informacijskega potenciala $V_R(\mathbf{E})$, podanega v (9).

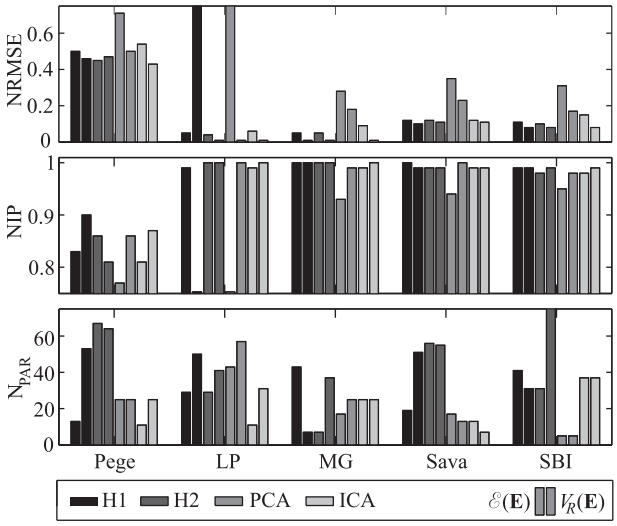
Pri minimizaciji povprečne kvadratne napake z gradientnimi metodami gradiente izračunamo po znanem vzvratnem postopku [17]. Po analognem postopku je mogoče izračunati tudi gradiente informacijskega potenciala [3]. Informacijski potencial ni odvisen od povprečja porazdelitve napak, zato se lahko zgodi, da po končani optimizaciji povprečje napak ne bo nič. Ker pa so izhodni nevroni linearni, lahko anomalijo odpravimo tako, da po končanem učenju ustrezno nastavimo pragove izhodnih nevronov [7].

5 Eksperimentalno delo

Omejili se bomo na modeliranje diskretnih časovnih vrst, vzorčenih v enakomernih časovnih presledkih. Za napovedovanje smo izbrali pet časovnih vrst: povprečno letno število sončevih peg (Pege), kaotično logistično preslikavo (LP), kaotično časovno vrsto Mackey-Glass (MG) ter tečaj delnice podjetja Sava (Sava) in vrednosti borznega indeksa SBI20 (SBI) v obdobju od začetka aprila 2007 do konca junija 2008. Modelirali smo na dva načina: pri prvem smo napovedovali vrednost časovne vrste v naslednjem koraku, pri drugem pa smo pričakovane spremembe vrednosti v časovni vrsti uvrščali v pet razredov. V vsaki časovni vrsti je bilo 308 podatkov, od katerih smo jih prvih 80 % uporabili za modeliranje, zadnjih 20 % pa za testiranje in primerjavo uspešnosti modelov. Da smo omejili velikost modelov, so le-ti smeli za napoved naslednje vrednosti uporabiti 12 predhodnih vrednosti, poleg tega pa število prostih parametrov modela ni smelo presegati 40 % števila podatkov v časovnih vrstah. Da bi se izognili naključnim zaustavitvam optimizacije v lokalnih minimumih, smo vsak model zgradili desetkrat, vsakič z naključno določenimi začetnimi parametri. V nadaljevanju so predstavljeni najboljši rezultati na testnih množicah.

5.1 Napovedovanje vrednosti

Za napovedovanje naslednje vrednosti v časovni vrsti z dvoplastnim perceptronom so bili uporabljeni štirje različni načini izbiranja značilk in dve kriterijski funkciji. Uspešnost napovedovanja je pri vsakem modelu podana s tremi cenilnimi funkcijami: s korenjeno povprečno kvadratno napako, normalizirano na standardno odstopanje časovne vrste od povprečja (*ang.* Normalized Root Mean Squared Error), $\mathrm{NRMSE} = \sqrt{\mathcal{E}(\mathbf{E})}/\sigma$, z normaliziranim informacijskim potencialom (*ang.* Normalized Information Potential), $\mathrm{NIP} = V_R(\mathbf{E})/\max\{V_R(\mathbf{E})\}$ in s številom prostih parametrov modela ($\mathrm{N_{PAR}}$). Rezultati so v obliki grafikonov predstavljeni na sliki 1. Na grafikonih vidimo, da sta vrednosti



Slika 1. Uspešnost napovedovanja naslednje vrednosti v časovni vrsti, ocenjena z različnimi cenilnimi funkcijami. Stolpci v enakem odtenku sive ustrezajo isti metodi iskanja značilk. Pri tem levi stolpec pomeni rezultat, dobljen pri modeliranju s kriterijsko funkcijo $\mathcal{E}(\mathrm{E})$, desni pa pri modeliranju s kriterijsko funkcijo $V_R(\mathbf{E})$
Figure 1. Quality of prediction of the future value in a time series estimated with different performance measures. Bars in the same shade of gray belong to the same preprocessing method, where the left and the right bar always represent results obtained by optimizing function $\mathcal{E}(\mathrm{E})$ and $V_R(\mathrm{E})$, respectively

NRMSE in NIP korelirani – majhna vrednost NRMSE se odraža v veliki vrednosti NIP in nasprotno.

Zaradi lažje primerjave metod so povprečne vrednosti uporabljenih cenilnih funkcij na vseh petih časovnih vrstah zbrane v tabeli 1. Pri večini časovnih vrst se je najbolje izkazalo izbiranje značilk s hevristično metodo

		H1	H2	PCA	ICA
	NRMSE	0,17	0,15	0,52	0,19
min	NIP	0,96	0,97	0,80	0,95
$\mathcal{E}(\mathbf{E})$	$\mathrm{N_{PAR}}$	29	38	21	19
max	NRMSE	0,32	0,14	0,22	0,13
$V_R(\mathbf{E})$	NIP	0,86	0,96	0,97	0,97
	$\mathrm{N_{PAR}}$	38	61	25	25

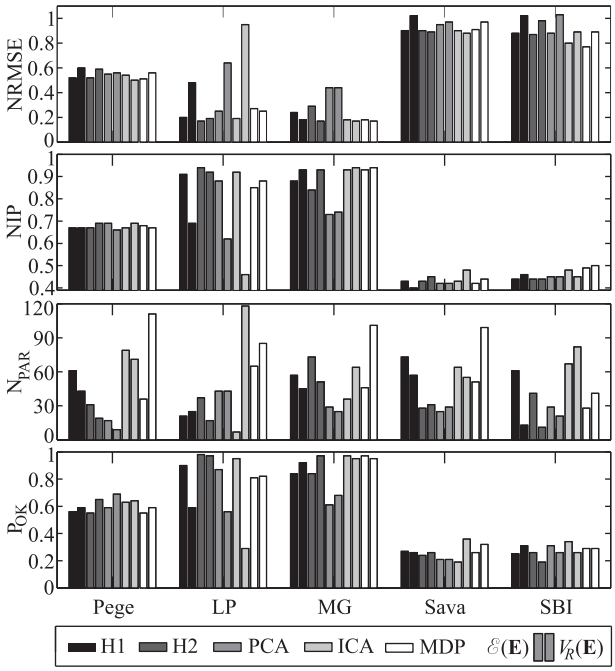
Tabela 1. Povprečne vrednosti cenilnih funkcij pri napovedovanju naslednje vrednosti v časovni vrsti
Table 1. Average values of performance measures (taken over all time series) in prediction of the future value in a time series

H2. žZal je ta metoda zaradi postopne gradnje nabora značilk na podlagi uspešnosti modela med vsemi računsko daleč najzahtevnejša. Po drugi strani daje izbiranje značilk z metodo PCA najslabše rezultate. Razlog verjetno tiči v tem, da manj pomembne glavne osi, ki se pri modeliranju ne uporabljajo, skrivajo pomembne podrobnosti o časovni vrsti. Pri modeliranju z minimizacijo povprečne kvadratne napake tudi metoda ICA ni prepričljiva. Nasprotno pa je pri modeliranju z maksimizacijo informacijskega potenciala izbiranje značilk z metodo ICA za malenkost uspešnejše kot z metodo H2, pri čemer imajo modeli v povprečju manj prostih parametrov. Poleg tega ta kombinacija izbiranja značilk in modeliranja v večini primerov pripelje do modelov, ki se izkazujejo z največjo vrednostjo NIP in najmanjšo vrednostjo NRMSE.

Na sliki 1 izstopajo visoke vrednosti NRMSE in s tem nizke vrednosti NIP pri napovedovanju logistične preslikave s hevrističnima metodama H1 in H2 v kombinaciji z maksimizacijo informacijskega potenciala. Kaotična logistična preslikava je podana z diferenčno enačbo, ki jo mora model za uspešno modeliranje rekonstruirati iz časovne vrste. To, da modeliranje le-te ni bilo uspešno v nobenem od 10 poskusov, nakazuje na slabše kovergenčne lastnosti pri optimizacije modelov z informacijskim potencialom.

5.2 Uvrščanje v razrede

V tem poskusu smo poskušali napovedati razliko med novo in zadnjo znano vrednostjo v časovni vrsti. Da bi lahko ovrednotili tudi metodo najznačilnejših projekcij, smo modele zasnovali tako, da pričakovano razliko uvrstijo v enega od petih razredov: veliko zmanjšanje, majhno zmanjšanje, neznatna sprememba, majhno povečanje in veliko povečanje. Meje med razredi so določene tako, da so le-ti kar najbolj enakovredno zastopani. Slika 2 prikazuje uspešnost uvrščanja napovedi v razrede za vseh pet časovnih vrst. Kot mero za uspešnost uvrščanja smo dodali še delež uvrstitev napovedi v pravi razred ($\mathrm{P_{OK}}$). Večje vrednosti NRMSE in manjše vrednosti NIP kažejo, da je tako spremenjeni problem težje modelirati. To še posebej velja za časovni vrsti Sava in SBI.



Slika 2. Uspešnost uvrščanja napovedi v razrede, ocenjena z različnimi cenilnimi funkcijami. Stolpci v enakem odtenku sive ustrezajo isti metodi iskanja značilk. Pri tem levi stolpec kaže rezultat, dobljen pri modeliranju s kriterijsko funkcijo $\mathcal{E}(\mathrm{E})$, desni pa pri modeliranju s kriterijsko funkcijo $V_R(\mathbf{E})$
Figure 2. Quality of classification of the predicted value in a time series estimated with different performance measures. Bars in the same shade of gray belong to the same preprocessing method, where the left and the right bar always represent results obtained by optimizing function $\mathcal{E}(\mathrm{E})$ and $V_R(\mathrm{E})$, respectively

		H1	H2	PCA	ICA	MDP
	NRMSE	0,55	0,55	0,61	0,52	0,53
min.	NIP	0,67	0,66	0,63	0,69	0,67
$\mathcal{E}(\mathbf{E})$	$\mathrm{N_{PAR}}$	55	42	29	51	45
	$\mathrm{P_{OK}}$	0,56	0,57	0,52	0,62	0,58
	NRMSE	0,66	0,56	0,73	0,68	0,57
max.	NIP	0,63	0,69	0,58	0,60	0,69
$V_R(\mathbf{E})$	$\mathrm{N_{PAR}}$	37	26	25	78	87
	$\mathrm{P_{OK}}$	0,53	0,61	0,48	0,50	0,60

Tabela 2. Povprečne vrednosti cenilnih funkcij pri uvrščanju napovedi v razrede
Table 2. Average values of the performance measures (taken over all time series) in classification of the predicted value to the five predefined classes

Zaradi lažje primerjave metod so v v tabeli 2 predstavljene povprečne vrednosti cenilnih funkcij čez vseh pet časovnih vrst. Pri optimizaciji povprečne kvadratne napake so v povprečju metode H2, ICA in MDP zelo

enakovredne. Če odmislimo časovno izredno zahtevno metodo H2, daje v tem primeru najboljše rezultate izbiranje značilk z metodo ICA. Prednost pred metodo MDP si je pridobila ravno z dobrim modeliranjem logistične preslikave. Pri optimizaciji informacijskega potenciala so rezultati večinoma najboljši v primerih, ko se značilke določajo z metodama H2 in MDP. Medsebojna primerjava uspešnosti modelov, dobljenih z optimizacijo povprečne kvadratne napake na eni strani in informacijskega potenciala na drugi, pokaže, da dajejo pri napovedovanju v razrede modeli, zgrajeni z optimizacijo povprečne kvadratne napake, boljše rezultate. Razlog je verjetno v tem, da je konvergenca proti optimumu informacijskega potenciala bistveno počasnejša kot proti optimumu povprečne kvadratne napake, zato se prva optimizacija pogosto konča v lokalnem minimumu.

Podobno kot pri napovedovanju vrednosti pri kaotični logistični preslikavi ponovno opazimo slabe rezultate, dobljene pri optimiziranju informacijskega potenciala (slika 2), kar ponovno kaže na slabše konvergenčne lastnosti pri uporabi te kriterijske funkcije.

6 Sklep

V prispevku so predstavljene možnosti uporabe mer, ki izhajajo iz teorije informacij, pri iskanju značilk v podatkih in pri samem modeliranju dinamičnih sistemov. Mere, ki izhajajo iz teorije informacij, so primerjane z nekaterimi klasičnimi merami pri gradnji dinamičnih modelov petih časovnih vrst. Rezultati so pokazali, da sta obe metodi iskanja značilk, ki izhajata iz informacijske teorije, boljši ali vsaj primerljivi z računsko intenzivnimi klasičnimi metodami. Slabše se je kot nadomestek povprečne kvadratne napake pri določanju prostih parametrov modela izkazal informacijski potencial. Poleg počasnega izračunavanja potenciala zaradi dvojne vsote v sami definiciji je konvergenca pri njegovi uporabi veliko počasnejša, zato se pogosto zgodi, da se optimizacija ustavi v lokalnem minimumu. Zaradi počasnejše konvergence je informacijski potencial uporabna mera predvsem v bližini optimuma, kjer bi lahko zamenjal klasično povprečno kvadratno napako.

7 Literatura

[1] A. Dobnikar, *Modeliranje nelinearnih dinamičnih sistemov na osnovi teorije informacij*, Znanje za trajnostni razvoj: zbornik povzetkov referatov 27. mednarodne konference o razvoju organizacijskih znanosti, Slovenija, Portorož, 32–45, 2008.

[2] J.C.A. van der Lubbe, *Information Theory*, Cambridge, Cambridge University, 1997.

[3] D. Erdogmus, J.C. Principe, *An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems*, IEEE trasactions on signal processing, 50, 1780–1786, 2002.

[4] F.C. Richards, T.P. Meyer, N.H. Packard, *Extracting cellular automation rules directly from experimental data*, Physica D, 45, 189–202, 1990.

[5] K.E. Hild, D. Erdogmus, J.C. Principe, *Blind source separation using Renyi's Mutual Information*, IEEE Signal Processing Letters, 8, 2001.

[6] U. Lotrič, A. Dobnikar, *Matrix formulation of the multilayered perceptron with a denoising unit*, Elektrotehniški vestnik, 70, 4, 221–226, 2003.

[7] D. Erdogmus, J.C. Principe, From Adaptive Filtering to Nonlinear Information Processing, *IEEE Signal Processing Magazine*, 23, 6, 14–33, 2006.

[8] J. Beirlant, E. Dudewicz, L. Gyorfi, E. van der Meulen, *Nonparametric entropy estimation: an overview*, International Journal of Mathematical and Statistical Sciences, 80, 1, 17-39, 1997.

[9] J.M. Santos, J.M. de Sa, L.A. Alexandre, *LEGClustA Clustering Algorithm Based on Layered Entropic Subgraphs*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30, 1, 62–75, 2008.

[10] A.A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*, Berlin Heidelberg, Springer, 2002.

[11] A. Gorban, B. Kegl, D. Wunsch, A. Zinovyev, *Principal Manifolds for Data Visualisation and Dimension Reduction*, New York, Springer, 2007.

[12] P. Comon, *Independent Component Analysis: a new concept?*, Signal Processing, 36, 3, 287–314, 1994.

[13] J.-F. Cardoso, *High-order contrasts for independent component analysis*, Neural Computation, 11, 157–192, 1999.

[14] A. Hyvarinen, *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis*, IEEE Transactions on Neural Networks, 10, 3, 626–634, 1999.

[15] J.-M. Wu, M.-H. Chen, Z.-H. Lin, *Independent component analysis based on marginal density estimation using weighted Parzen windows*, Neural Networks, 21, 7, 914–924, 2008.

[16] K.E. Hild, D. Erdogmus, K. Torkkola, J.C. Principe, *Sequential Feature Extraction Using Information Theoretic Learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 9, 1385–1393, 2006.

[17] S. Haykin, *Neural Networks: A Comprehensive Foundation*, New Jersey, Prentice-Hall, 1999.

Marko Bratina je leta 2006 magistriral na Univerzi v Ljubljani, Fakulteti za elektrotehniko. Zaposlen je v podjetju Savatech v Kranju kot vodja elektroprojektive in vzdrževanja elektronike. Dela na področju procesne avtomatizacije, v zadnjem času tudi pri uvajanju adaptivnih sistemov v proizvodni proces.

Andrej Dobnikar je redni profesor na Univerzi v Ljubljani, Fakulteti za računalništvo in informatiko. Raziskovalno se ukvarja z metodami mehkega računanja, porazdeljenimi in adaptivnimi sistemi.

Uroš Lotrič je docent na Univerzi v Ljubljani, Fakulteti za računalništvo in informatiko. Raziskovalno dela na področjih nevronskih mrež v povezavi z informacijsko teorijo in porazdeljenim procesiranjem.