UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VPLYV TOPOLÓGIE SKRYTEJ VRSTVY ECHO SIETÍ NA KVALITU VYGENEROVANÝCH REGRESOROV

DIPLOMOVÁ PRÁCA

2015 Bc. Róbert TÓTH

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VPLYV TOPOLÓGIE SKRYTEJ VRSTVY ECHO SIETÍ NA KVALITU VYGENEROVANÝCH REGRESOROV

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	9.1.9. Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Dr. Eng. Ján Dolinský

Bratislava 2015 **Bc. Róbert TÓTH**

74453042

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Róbert Tóth
Študijný program:	ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	9.1.9. aplikovaná matematika
Typ záverečnej práce:	diplomová
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Vplyv topológie skrytej vrstvy echo sietí na kvalitu vygenerovaných regresorov
On a relationship between the topology of echo state networks and quality of echo regressors

Cieľ: Cieľom práce je oboznámiť sa s problematikou echo sietí a grafových štruktúr používaných pri tvorbe ich topológii. Z hľadiska čísla podmienenosti a strednej kvadratickej chyby otestujeme vplyv rôznych topológii na kvalitu sieťou vygenerovaných regresorov. Testovať budeme na syntetických aj reálnych dátach.

Vedúci:	Dr. Ing. Ján Dolinský
Katedra:	FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry:	prof. RNDr. Daniel Ševčovič, CSc.
Dátum zadania:	29.01.2014

Dátum schválenia: 10.02.2014 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou sa chcem poďakovať svojmu vedúcemu diplomovej práce Dr. Eng. Jánovi Dolinskému za ochotu, pomoc, odborné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce. Ďakujem aj svojim priateľom a rodine za pomoc pri tvorbe obrázkov a jazykovú a štylistickú úpravu textu.

Abstrakt v štátnom jazyku

Vplyv topológie skrytej vrstvy echo sietí na kvalitu vygenerovaných regresorov [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Dr. Eng. Ján Dolinský, Bratislava, 2015

V našej práci skúmame vlastnosti echo sietí. Meníme topológiu ich skrytých vrstiev a pozorujeme kvalitatívne zmeny v zásobníku regresorov vygenerovanom sieťou pomocou čísla podmienenosti. Na tieto účely používame rozličné konfigurácie riedkych, náhodných, scale free a small world sietí. Náš záver je, že s rastúcim priemerným stupňom neurónov v skrytej vrstve číslo podmienenosti zásobníka klesá. Zisťujeme, že má zmysel hrany v sieti ovážiť aj kladnými aj zápornými znamienkami a že použitím alternatívneho rozdelenia namiesto rovnomerného dosiahneme rýchlejšiu konvergenciu k minimálnemu číslu podmienenosti. Navrhujeme aj niekoľko možných dôvodov tohoto správania - neuróny s malou konektivitou sú zrejme zdrojom multikolinearity.

Kľúčové slová: echo siete, scale-free siete, small-world siete, číslo podmienenosti

Abstract

On a relationship between the topology of echo state networks and quality of echo regressors [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Dr. Eng. Ján Dolinský, Bratislava, 2015

In our work we investigate the properties of the echo state networks. We change the topologies of their hidden layers and observe the qualitative changes in reservoir consisting of generated regressors via conditional number. For these purposes we use different configurations of sparse, completely random, scale free and small world topologies. We conclude that with higher mean degrees of neurons in the hidden layer we can expect lower conditional numbers. We find out, that there is a reason for using both positive and negative weights for connections and with the alternative distribution of weights we obtain faster convergence to the minimal conditonal number than with the uniform distribution. We propose couple of possible reasons for this behavior. Neurons with small connectivity might be a source of multicolinearity.

Keywords: echo state network, scale-free network, small-world network, conditional number

Obsah

Úvod

Echo state siete sú v posledných rokoch ostro sledovanou oblasťou strojového učenia, pretože napriek tomu, že stoja na jednoduchej myšlienke a sú výpočtovo veľmi nenáročné, majú výnimočnú schopnosť generalizácie. V algoritme sa snúbili dobré vlastnosti lineárnej regresie a neurónových sietí, čo umožňuje nastaviť mnoho parametrov tak, aby sieť performovala čo najlepšie. Zároveň sa algoritmus dá obohatiť o mnoho prístupov vlastných pre samotnú lineárnu regresiu, ako sú dopredný výber a 'hrebeňová' regularizácia. Čo je na jednej strane výhodou, je na strane druhej zároveň aj výzvou - ako nastaviť jednotlivé parametre a postupy tak, aby echo sieť dosahovala dobré výsledky? V prvej kapitole sa venujeme uvedeniu do problematiky a základným vlastnostiam.

Tak ako každý learning algoritmus, tak aj tento sa musí popasovať s problémom multikolinearity, ktorému sa venujeme v druhej kapitole. Pozrieme sa na to, ako vzniká a aké sú rozličné prístupy pri jeho adresovaní. Jedným z nich je meniť štruktúru skrytej vsrtvy z náhodnej na konkrétne grafové topológie. Preto sa v tretej kapitole pozrieme na scale free a small world siete a vysvetlíme si, ako sa vytvárajú.

Vo štvrtej kapitole sa venujeme číslu podmienenosti matice, ktoré budeme používať ako hlavné merítko pri určovaní kvality zásobníka regresorov, ktoré sieť vygeneruje pred aplikovaním samotnej lineárnej regresie. Následne v piatej kapitole otestujeme rôzne topológie a ich parametre. Odpozorované javy sa pokúsime objasniť v kapitole šiestej. Na záver sa venujeme predikčnej schopnosti modelu a ďalším možnostiam použitia nami dosiahnutých výsledkov.

1 Echo state network

Echo state network (ďalej len ESN) je algoritmom strojového učenia, ktorý spája dobré vlastnosti lineárnej regresie a neurónových sietí. Na rozdiel od klasickej neurónovej siete však skrytá vrstva neobsahuje tradičné perceptróny s aktivačnou sigmoidálnou funkciou, ale perceptróny schopné pamätať si jednotlivé stavy v nich, čo ju robí obzvlášť vhodnou na predikcie vývoja časových radov. Takáto trieda sietí sa zvykne označovať ako rekurentné neurónové siete - RSN. Dôležitý rozdiel medzi NS a RNS je nemennosť váh siete v skrytej vrstve počas priebehu učenia - kým v NS sa váhy medzi neurónmi menia každou iteráciou vstupov, t.j. k učeniu dochádza v každom kroku, ESN sa 'učí' iba raz, a to v úplne poslednom kroku, keď boli všetky vstupy vo všetkých časoch už zadané. Práve preto sa ESN ako learning algoritmus ponáša oveľa viac na lineárnu regresiu ako na samotné neuronové siete. Jeho obrovskou výhodou však je, že kým obyčajná lineárna regresia má na porúdzi iba čisté prediktory vo forme vstupov, ESN vytvorí mnoho ďalších umelých prediktorov, ktoré v sebe nesú zaujímavé časovo-priestorové interakcie tých pôvodných. Tie môžu skrývať veľa inak nevyužitej informácie.

Ako príklad podporujúci toto tvrdenie môže poslúžiť nasledovná situácia. Predstavme si, že máme dva prediktory - jeden popisujúci víkendovú informáciu, t.j. schodovitý signál s hodnotou 1 cez víkend a 0 cez týždeň, a teplotu. Ako výstup sa snažíme modelovať množstvo spotrebovanej elektriny v meste. Je zjavné, že ak naše dva prediktory vynásobíme, dostaneme úplne novú, ale stále užitočnú informáciu, pretože ľudia sa cez víkendy správajú značne odlišne: ak je počasie pekné (vysoká teplota) - idú von na prechádzku a nespotrebujú takmer nič a ak je škaredé (nízka teplota) - ostanú doma a používajú spotrebiče. Tak isto má zmysel uvažovať aj posunutia prediktorov v čase. Je zjavné, že spotreba elektriny v utorok o šiestej bude veľmi podobná tej v pondelok o tom istom čase. ESN všetky takéto časovo-priestorové expanzie prediktorov dokáže vytvoriť. Pozrime sa najprv na základné definície.

1.1 Definícia a základné vlastnosti

ESN zadefinujeme ako v pôvodnom článku [4]:

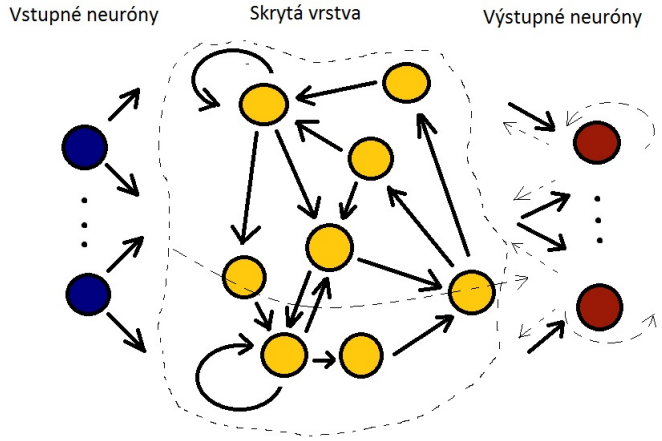
Definícia 1.1. *Uvažujme neurónovú sieť s diskrétnym časom, K vstupmi, N neurónmi v skrytej vrstve a L výstupmi. Vstupy v časovom kroku n sú $u(n) = u_1(n), \cdots, u_K(n)$, hodnota v neurónoch v skrytej vrstve $x(n) = x_1(n), \cdots, x_N(n)$ a $y(n) = y_1(n), \cdots, y_L(n)$ vo výstupnej vrstve. Váhy medzi vstupnou a skrytou vrstvou (v tomto smere) sú reálne čísla uchovávané v matici W^{in} rozmerov $N \times K$, matica W rozmerov $N \times N$ obsahuje váhy skrytej vrstvy, matica W^{out} rozmerov $N \times (K+N+L)$ obsahuje váhy spojení medzi všetkými vrstvami a výstupnou vrstvou (v tomto smere) a nakoniec matica W^{back} rozmerov $N \times L$ obsahuje váhy spojení medzi výstupnou vrstvou a skrytou (v tomto smere). Neuróny skrytej vrstvy aktualizujeme na základe vzťahu*

$$x(n+1) = f(W^{in}u(n+1) + Wx(n) + W^{back}y(n)),$$

kde $f = (f_1, \cdots, f_N)$ sú typicky sigmoidálne funkcie. Predikovaný výstup potom počítame ako

$$y(n+1) = f^{out}(W^{out}(u(n+1), x(n+1), y(n))),$$

kde $f^{out} = (f_1^{out}, \cdots, f_N^{out})$ a $(u(n+1), x(n+1), y(n))$ je spojením hodnôt príslušných vektorov.



Obr. 1: všeobecná schéma ESN - v tej našej sa čiarkované hrany používať nebudú

V tejto diplomovej práci sa obmedzíme na špeciálny prípad ESN, keď f^{out} je identita, f hyperbolický tangens, W^{back} je nulová matica a výstupná vrstva neinteraguje sama so sebou. Posledné dva vzťahy definície sa potom zmenia na

$$x(n+1) = \tanh(W^{in}u(n+1) + Wx(n)) = E(x(n), u(n+1)) \tag{1}$$

a

$$y(n+1) = W^{out}(u(n+1), x(n+1)). \tag{2}$$

Môžeme si všimnúť, z čoho plynie príznačný názov 'echo' sieť. Stav každého neurónu v čase nesie informáciu o vstupoch nielen z daného času, ale 'doznievajú' v ňom aj stavy z minulosti, ktoré boli ovplyvnené minulou hodnotou vstupov. V tomto bode by sme si ale mali uvedomiť, že takto nastavené podmienky siete môžu spôsobiť, že jednotlivé stavy v čase budú značne odlišné pre rôzny počiatočný stav v čase nula, ktorý však je daný vstupmi, nie nami. Takéto správanie modelu je samozrejme nežiaduce, pretože je s ním spätá aj vysoká nestabilita pri predpovediach a jeho natrénovanie by potrebovalo mnoho pokusov. Preto sa od ESN vyžaduje aj takzvaná 'vlastnosť echo stavov', ktorej cieľom je zabezpečiť, aby naozaj platilo, že

$$x(n) = f(\cdots, u(n-1), u(n))$$

a počiatočný stav $x(0)$ v modeli po čase nezohrával žiadnu úlohu.

Definícia 1.2. *Sieť má echo stavy, ak je stav $x(n)$ jednoznačne určený zľava nekonečnou postupnosťou vstupov $\cdots, u(n-1), u(n)$. Presnejšie, pre každú takúto zľava nekonečnú postupnosť a dve zľava nekonečné postupnosti $\cdots, x(n-1), x(n)$ a $\cdots, x'(n-1), x'(n)$, kde $x(i) = E(x(n-1), u(i))$ a $x'(i) = E(x'(n-1), u(i))$, platí, že $x(n) = x(n-1)$.*

V realite však potrebujeme túto vlastnosť zabezpečiť už pri konštrukcii siete. Je zrejmé, že bude záležať hlavne na vlastnostiach matíc W a W^{in}. Dôležitým výsledkom je preto pre nás nasledujúca veta.

Veta 1.3. *Nech je najväčšia vlastná hodnota matice W rovná $\lambda_{max} = \Lambda < 1$. Potom platí, že $d(E(x,u), E(x',u)) < \Lambda d(x,x')$ pre všetky vstupy u a stavy $x, x' \in (-1,1)^N$, kde d značí euklidovskú vzdialenosť. ESN má preto echo stavy.*

Dôkaz. Dôkaz vety môžeme nájsť v apendixe [4]. □

Môžeme si všimnúť hneď niekoľko zaujímavých vecí. Veta nekladie žiadne podmienky na maticu W^{in}, čo zrejme dáva zmysel, pretože k samotnému iterovaniu používania minulých stavov dochádza prostredníctvom matice W. Ďalej nám veta dáva

jednoduchý návod ako takúto sieť vyrobiť z už existujúcej štruktúry W, ktorá môže a nemusí túto vlasnosť zabezpečovať. Nech λ_{max} označuje v absolútnej hodnote najväčšiu singulárnu hodnotu matice W. Z algebry vieme, že platí vzťah $\lambda_{max}(\alpha W) = \alpha\lambda_{max}(W)$. Ak teda stanovíme $\alpha_{min} = 1/\lambda_{max}$, pre všetky hodnoty α menšie ako α_{min} bude mať ESN s αW definujúcou skrytú vsrtvu echo stavy. V praxi sa ukazuje, že táto podmienka je síce postačujúca, ale príliš prísna a hranica α_{max}, nad ktorou už echo stavy zaručene mať nebudeme, je o čosi vyššie. [7] [5]

Našli sme teda návod, ako generovať sieť s echo stavmi.

1. Zvolíme si počet neurónov v skrytej vrstve (v definícii N) a na základe neho vygenerujeme matice W^{in} a W.

2. Matica W^{in} sa obyčajne po prvkoch generuje z rovnomerného rozdelenia na $(-1, 1)$.

3. Maticu W môžeme charakterizovať jej topológiou (t.j. umiestnením nenulových prvkov) a následne prenásobiť maticou totožných rozmerov s prvkami z rovnomerného rozdelenia na $(-1, -1)$. Takto dosiahneme ováženie jednotlivých spojení medzi neurónmi.

4. Maticu W vhodne preškálujeme, aby sme dosiahli echo stavy - napr. vydelíme najväčšou vlastnou hodnotou.

5. Spustíme sieť a postupne do nej feedujeme hodnoty $u(n)$. Na základe vzťahu (1) aktualizujeme hodnoty neurónov $x(n)$ v skrytej vrstve.

6. Zozbierame jednotlivé časové rady, ktoré vznikli v neurónoch ako postupnosti ich stavov. Tie dokopy po stĺpcoch vytvárajú maticu plánu, ktorú nazveme X'. Ak časové rady jednotlivých postupností vstupov po stĺpcoch vložíme do matice U, vidíme, že nájsť výstupný vektor W^{out} je klasickou úlohou lineárnej regresie

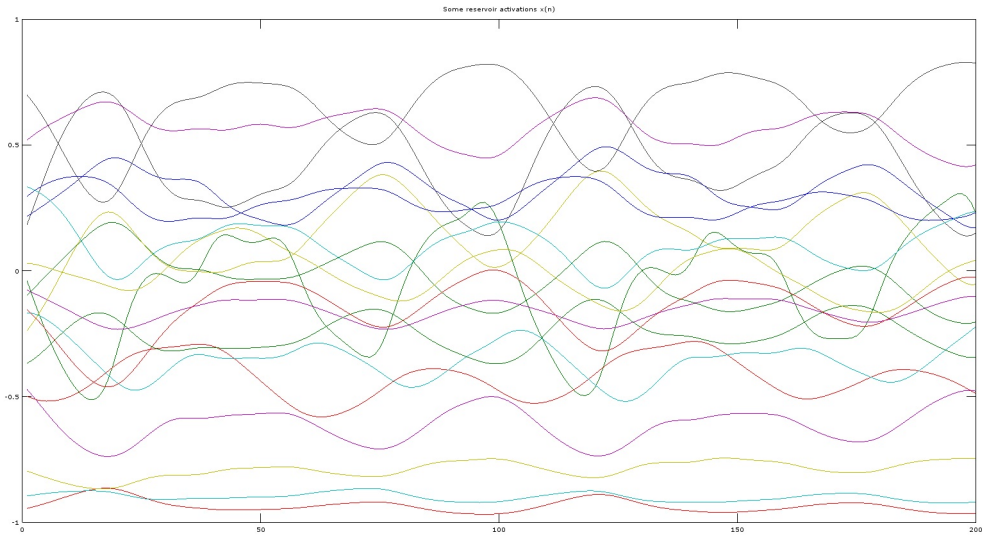
$$Y = U\beta_u + X'\beta_x + \varepsilon,$$

pričom $W^{out} = (\beta_u, \beta_x)$.

Pre jednoduchšiu notáciu spojíme matice X' a U do matice X, čím sa náš vzťah zmení na

$$Y = XW^{out} + \varepsilon.$$

Na maticu X sa v ďalšom texte budeme odvolávať ako na ‘zásobník‘. Po skonštruovaní zásobníka by sme mali zahodiť nejakú jeho začiatočnú časť, ktorá je ovplyvnená stavom $x(0)$. Ak bola matica W skonštruovaná dobre, ESN by mala na tento stav ‘zabudnúť‘ už po niekoľkých desiatkach časových iterácií. [4] V našich pokusoch sme zahadzovali prvých 100 riadkov konečného zásobníka.

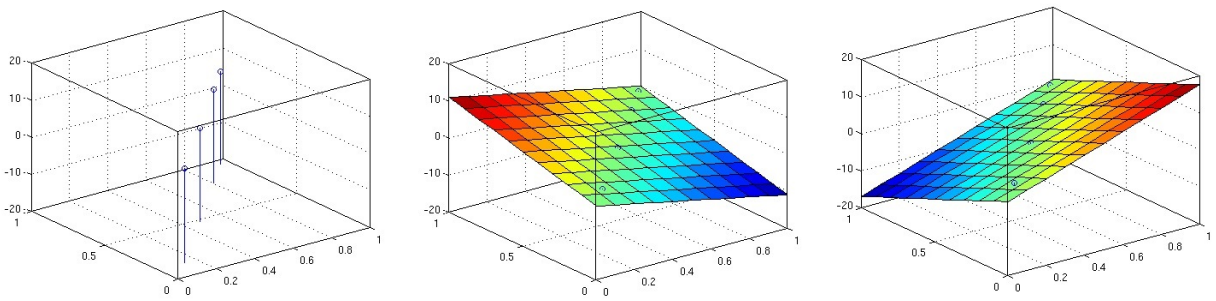


Obr. 2: niektoré zo signálov v zásobníku

Naším cieľom bude pozrieť sa na to, ako topológia matice W vplýva na kvalitu regresorov vytvorených touto procedúrou.

2 Zdroje kolinearity v ESN

Veľkým problémom, s ktorým sa musí lineárna regresia popasovať, býva multikolinearita prediktorov. Keďže riešenie sa hľadá v tvare $(X^\top X)^{-1}X^\top Y$, je nutné, aby matica $X^\top X$ mala dobré numerické vlastnosti. Ak sú však vektory v X kolineárne, táto vlastnosť pre ňu samotnú neplatí. V matici $(X^\top X)^{-1}$ sa táto numerická nestabilita ešte znásobí, pretože, ako si ukážeme v štvrtej kapitole, jej číslo podmienenosti vzrastie kvadraticky. Preto je cieľom vytvoriť taký zásobník X, ktorý by mal vektory dobre dekorelované. Ďalším dôvodom, prečo udržať model s čo najmenej kolineárnymi prediktormi, je stabilita predikcií. Malá odchýlka v nameraných hodnotách mohla totiž pri trénovaní modelu spôsobiť, že model aproximuje funkčnú závislosť úplne zle, ako demonštruje fitovanie nadrovinou na nasledujúcom obrázku.

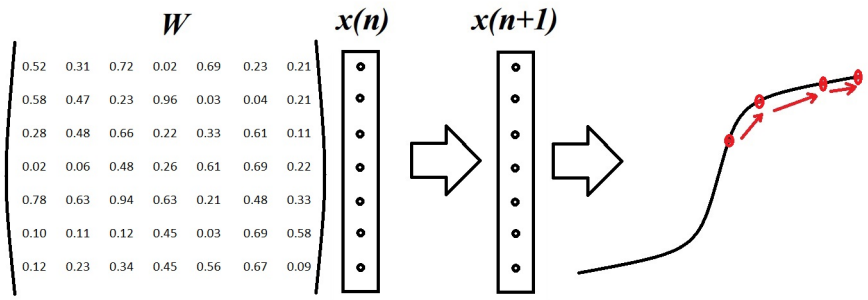


Obr. 3: nestabilita modelu vytvorenom na kolineárnych prediktoroch

Pozrime sa teraz na jeden z častých dôvodov, prečo ku korelovaniu vektorov v matici X dochádza a ako sa s ňou dá vyrovnať.

Zoberme si ESN, ako ju máme zadefinovanú a predstavme si, že by matica W obsahovala iba čísla generované z rovnomerného rozdelenia na intervale $(0, 1)$. Na chvíľu zanedbajme vplyv prediktorov u v n-tom kroku kroku iterácie a tvárme sa, že do aktivačnej sigmoidálnej funkcie vstupuje iba vektor $Wx(n)$. Ak sa niekedy stav vektora $x(n)$ dostane do bodu, že sú jeho zložky kladné, je jasné, že v ňom ostane aj naďalej. Po vstupe do aktivačnej funkcie - v našom prípade hyperbolický tangens - ostane kladným tiež. Vďaka prenásobovaniu maticou W v každom časovom kroku sa však hodnoty vstupujúce do sigmoidy budú neustále zvyšovať, až kým sa všetky úplne nepriblížia k jednotke. Po nejakom čase preto budú stavy v jednotlivých neruónoch takmer nerozlí-

šiteľné, čo spôsobí, že jednotlivé vektory zásobníka budú veľmi korelovať. Analogicky to platí aj pre záporné znamienka vo W, prípadne pre nešťastnú kombináciu alokácie tých kladných a záporných.

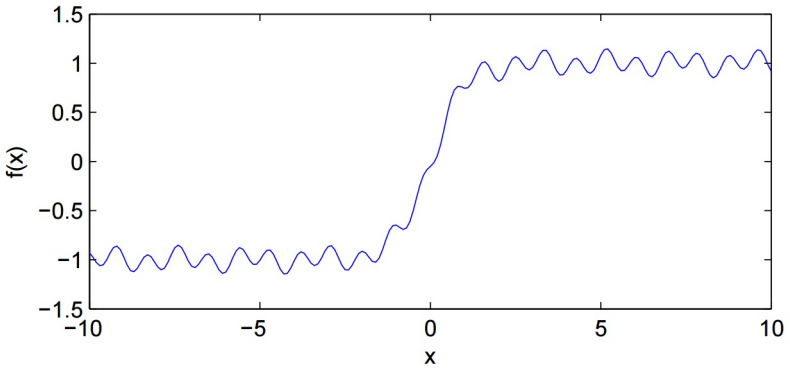


Obr. 4: postupné korelovanie vektorov v čase

Problém multikolinearity sa v praxi ľudia snažili adresovať rôznymi prístupmi. Jedným z nich je napríklad obohatiť sigmoidu o periodické výrazy

$$f(x) = \tanh(x) + \omega_1 sin(p_1 x) + \omega_2 sin(p_2 x),$$

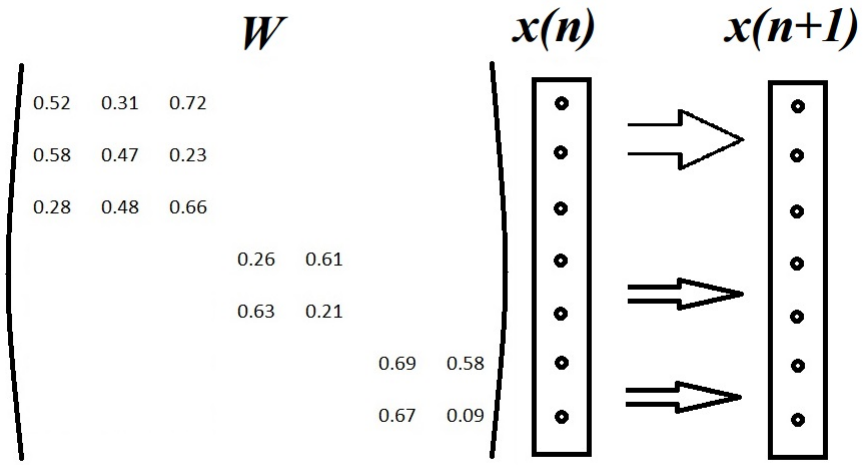
čo umožní rozlíšiť signály, ktoré by sme v nami uvedenom príklade rozlíšiť nevedeli. [3] Problém s aktivačnou funkciou je však len jeden z mnohých, ktoré multikolinearitu môžu spôsobiť, preto hľadanie rozličných prístupov, ako ju adresovať, je nutnou úlohou štúdia ESN.



Obr. 5: sigmoidálna funkcia obohatená o periodické výrazy

Mnoho autorov a publikácií sa snaží pozrieť na tento problém z hľadiska topológie matice W. [7] Predstavme si na chvíľu, že matica W je bloková. Potom je jasne vidieť,

že jednotlivé časti vektora x(n) ‘žijú‘ samostatným životom, pretože nedostávajú informáciu o ostatných, keďže medzi danými neurónmi neexistujú spojenia. Takýto prístup nám potom môže pomôcť zaručiť dekorelovanosť jednotlivých neurónov naprieč zásobníkom. Na uvedenej myšlienke stojí štúdium tzv. ‘small-world‘ sietí, ktoré spĺňajú požadovanú vlastnosť zhlukovania jednotlivých neurónov, a preto boli vhodným kandidátom na tvorbu topológií ESN.



Obr. 6: zhluky v matici W a ich vplyv na dekorelovanie

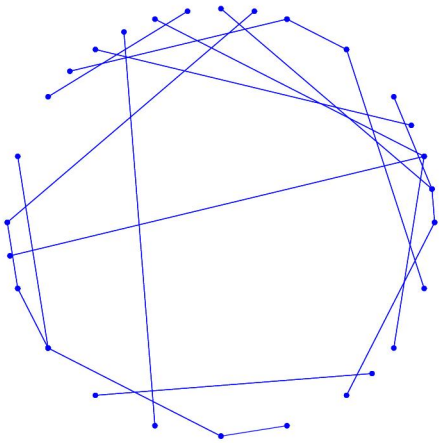
Tento prístup budeme skúmať bližšie aj my. Predtým ale potrebujeme zadefinovať rozličné typy najčastejšie používaných topológií ESN.

3 Topológie sietí

3.1 Pôvodná topológia

Jaeger v úvodnom článku uvádzajúcom ESN navrhuje používať topológiu matice W, ktorá je veľmi riedka - pomer nenulových prvkov oproti všetkým je len okolo päť percent. [4] Neskoršie pokusy s ESN ukázali, že držať maticu riedku z hľadiska výkonu a udržania echo stavov nemá až taký veľký význam [5], ale mať tento pojem zadefinovaný je napriek tomu užitočné. Pojem topológia siete je ničím iným ako grafom s nejakými vlastnosťami. Jeho matica susednosti je tak to isté, ako matica W (v prípade, že hrany dodatočne neovážime).

Definícia 3.1. *Konektivitou v grafe G budeme nazývať pomer počtu jeho hrán k počtu hrán, ktorý by mal kompletný graf na takom istom počet vrcholov. Ak je daná matica susednosti, ide jednoducho o pomer počtu jej nenulových prvkov k všetkým. Pre graf na n vrcholoch tak dostávame $2 * \|E\|/(n(n-1))$, kde E označuje množinu jeho hrán.*

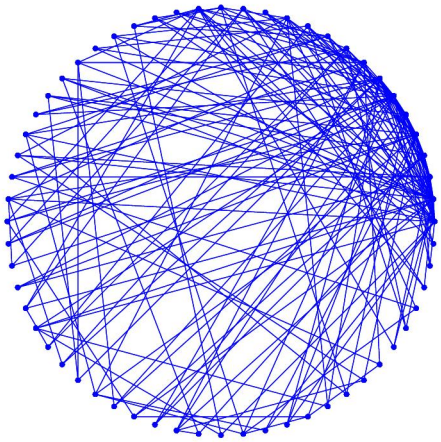


Obr. 7: pôvodná topológia s konektivitou 5 percent

3.2 Scale-free siete

Scale-free sieť (ďalej len SF) je druh siete, v ktorej počet vrcholov s daným stupňom klesá exponenciálne s veľkosťou stupňa. Presnejšie povedané, ak $P(k)$ označuje počet

vrcholov v grafe stupňa k, potom $P(k) \sim k^{-\gamma}$, kde γ sa obyčajne pohybuje medzi hodnotami 2 až 3. Takéto siete sa bežne vyskytujú v každodennom svete, či už je to sieť citácií v akademickej obci, alebo internet. [1] Ich výskum sa preto samozrejme dotkol aj problematiky ESN. Algoritmov na ich tvorbu je niekoľko. Uvedieme najstarší z nich, ktorý sme použili aj pri našich simuláciách v neskorších kapitolách.



Obr. 8: SF sieť - na jeho pravej strane vidíme veľký zhluk

Barabásiho–Albertov model

Algoritmus začneme vytvorením spojenej siete m_0 počiatočných vrcholov. Postupne do grafu po jednom pridávame nové vrcholy. Každý z týchto nových vrcholov je následne spojený s $m \leq m_0$ vrcholmi, pričom týchto m sa vyberá náhodne. Rozdelenie však nie je rovnomerné, ale vrcholom s veľkým stupňom prisúdime väčšiu pravdepodobnosť spojenia s novým - presnejšie platí, že

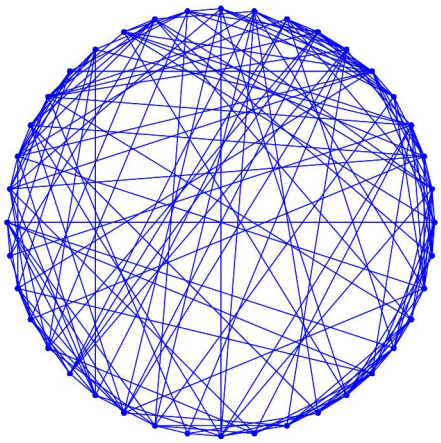
$$p_i = \frac{k_i}{\sum_j k_j},$$

kde p_i označuje pravdepodobnosť spojenia s i-tym vrcholom, k_i stupeň tohoto vrcholu a suma k_j počet všetkých hrán v grafe existujúcich v čase pridania nového vrcholu.

Keďže nové vrcholy majú tendenciu pridávať sa k hlavne veľkým uzlom, konečná štruktúra potom naozaj zachováva vlasnosť malého počtu vrcholov s veľkým stupňom a naopak, ‘osamotených‘ vrcholov je veľa.

3.3 Smal-world siete

Veľmi populárnou topológiou v ESN sú tzv. Small-world siete (ďalej len SW). Túto štruktúru môžeme nájsť v prírode aj vo svete ľudí - či už je to nervová sústava červa, elektrická sieť U.S.A. alebo konexie amerických hercov. Tieto siete sú charakterizované zhlukmi vrcholov, ktoré dokážeme nájsť naprieč celým grafom. Algoritmov na ich tvorbu je opäť niekoľko, ale my uvádzame ten z pôvodného článku, kde boli SW siete uvedené prvýkrát a ktorý sme použili aj pri našich simuláciách. [6] Jeho myšlienka stojí na pokuse nájsť interpoláciu medzi kompletne náhodnými grafmi a grafmi s pravidelnou štruktúrou.

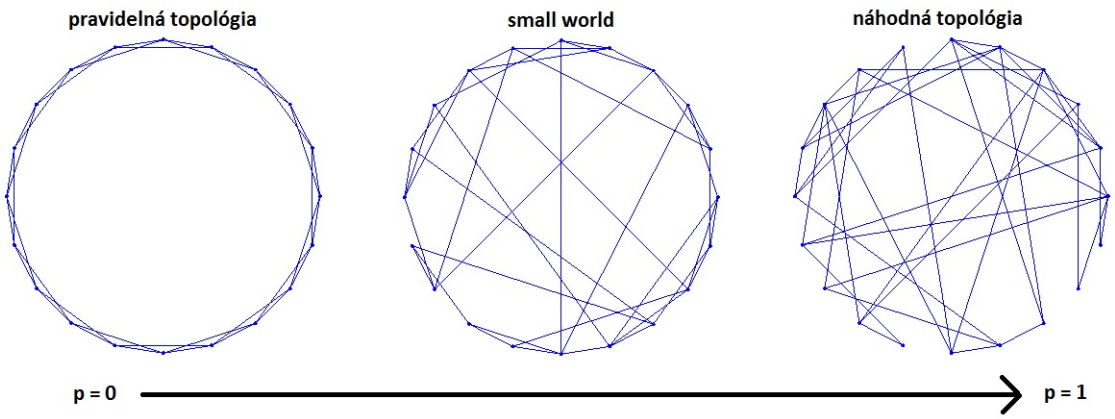


Obr. 9: SW sieť - môžeme si všimnúť zhruba 3 väčšie zhluky

Rewiring procedúra

Algoritmus začína pravidelným prstencovým grafom s n vrcholmi, pričom každý vrchol je spojený s k najbližšími. Vyberieme si vrchol a hranu, ktorá ho spája s jeho najbližším susedom vzhľadom na smer hodinových ručičiek a prstenec. S pravdepodobnosťou p túto hranu odstránime a pripojíme náhodne na miesto neexistujúcej hrany medzi dvoma vrcholmi. Túto hranu vyberáme náhodne rovnomerne cez všetky možné, pričom duplicitné hrany sú zakázané. Tento proces zopakujeme postupne pre všetky vrcholy v smere hodinových ručičiek. Keď dokončíme jedno kolo, celú procedúru zopakujeme analogicky pre druhé najbližšie vrcholy a ich hrany. Takto pokračujeme, až kým nevykonáme aspoň $k/2$ kôl. So zvyšujúcou sa pravdepodobnosťou p narastá v grafe

chaos, až kým nie je úplne náhodný. Pre stredné hodnoty p dostávame SW sieť.



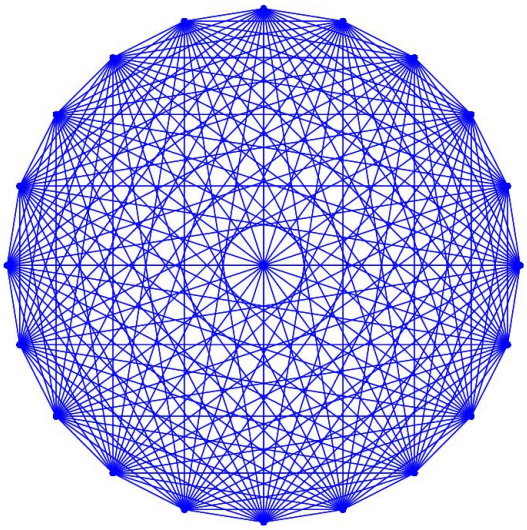
Obr. 10: interpolácia medzi pravidelným a kompletne náhodným grafom

Môžeme si všimnúť, že počet hrán v grafe zostáva počas celého algoritmu invariantný, a preto sa zachováva aj priemerný stupeň vrcholov (počiatočné k) a celková konektivita v grafe, ktorú môžeme vypočítať ako $(nk/2)/(n(n-1)/2) = k/(n-1)$. Tieto grafy sa väčšinou charakterizujú pomocou pojmov 'characteristic path length' a 'clustering coefficient', ale my sa na ne budeme pozerať skôr z konštrukčného hľadiska - teda na hodnoty n, k a p. Dá sa ukázať, že SF grafy sú špeciálnou triedou SW, ktorú nazývame aj 'ultra-SW' [2]. Autori odporúčajú hodnotu k nastaviť tak, aby $ln(n) << k << n$. Prvú nerovnosť zdôvodňujú tým, že pre menšie k sa môže graf rozpadnúť na viacero komponentov a dôvod pre druhú neuvádzajú. Zrejme však pôjde o výpočtovú zložitosť algoritmu.

3.4 Kompletný graf

Takáto topológia má tiež význam, pretože sa hrany medzi vrcholmi v ESN zvyknú ovážiť, čím vieme dosiahnuť variabilitu medzi na pohľad rovnakými topológiami. Táto sieť je zároveň výsledkom predchádzajúceho algoritmu na konštrukciu SW siete pre špeciálny prípad $k = n$.

Pozn. Vo všetkých grafoch sa môžu uvažovať aj slučky napriek tomu, že pre estetickosť na obrázkoch nie sú znázornené. My sme slučky uvažovali iba v poslednom prípade



Obr. 11: kompletný graf

kompletného grafu.

Teraz, keď sme si zadefinovali topológie sietí, ktorých vlastnosti chceme skúmať, potrebujeme si vytvoriť aj aparát na ohodnotenie kvality zásobníka X.

4 Kvalita zásobníka

4.1 Číslo podmienenosti matice

V literatúre sa zvyknú rozlišovať pojmy absolútne číslo podmienenosti a relatívne číslo podmienenosti. Keďže v našej diplomovej práci sa venujeme skôr konkrétnym aplikáciám ESN ako teoretickým vlastnostiam, obmedzíme sa len na druhý pojem a budeme ho označovať skrátene ako číslo podmienenosti.

Pri riešení systému lineárnych rovníc $Ax = b$ nie sme vždy schopní nájsť presné riešenie (ak b nepatrí do stĺpcového priestoru matice A), iba nejakú jeho v istom zmysle najlepšiu aproximáciu. V praxi sa častokrát stáva, že vektor b nedostaneme v čistej forme ale zašumený nejakou chybou. V ideálnom prípade by mala malá chyba v b spôsobiť iba malú chybu v riešení x. Ak tomu tak nie je, matica A mala zlé vlastnosti - hovoríme, že bola zle podmienená.

Označme chybu vo vektore b ako e. Ak existuje inverzná matica k matici A, chyba v riešení $A^{-1}b$ je $A^{-1}e$. Pomer relatívnej chyby riešenia k relatívnej chybe vo vektore b je potom

$$er = \frac{\|A^{-1}e\|/\|A^{-1}b\|}{\|e\|/\|b\|}.$$

Z vlastností maticových noriem vieme, že pre 'rozumné' normy (napr. indukované) platí $\|A^{-1}e\| \leq \|A^{-1}\|\|e\|$ a $b = \|AA^{-1}b\| \leq \|A\|\|A^{-1}e\|$. Z toho potom

$$er \leq \|A^{-1}\|\|A\| = c(A).$$

Znakom c(A) budeme označovať už spomínané číslo podmienenosti matice. Inverz matice A nemusí vždy existovať, preto je dobré mať k dispozícii aj alternatívnu definíciu.

Veta 4.1. *Pre euklidovskú normu matice $\|.\|_2$ platí*

$$c(A) = \frac{\sigma_{max}}{\sigma_{min}},$$

kde σ_{max} a σ_{min} označujú najväčšiu a najmenšiu singulárnu hodnota matice A. Singulárne hodnoty matice A dostaneme ako odmocniny vlastných hodnôt matice $A^\top A$.

Dôkaz. Z definície prepíšeme normu a použijeme singulárny rozklad:

$$\begin{aligned}
\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sup_{x\neq 0} \frac{\|U\Sigma V^H x\|_2}{\|x\|_2} \\
&= \sup_{x\neq 0} \frac{\|\Sigma V^H x\|_2}{\|x\|_2} \\
&= \sup_{x\neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2} \\
&= \sup_{x\neq 0} \frac{(\sum_{i=1}^{r} \sigma_i^2 |y_i|^2)^{\frac{1}{2}}}{(\sum_{i=1}^{r} |y_i|^2)^{\frac{1}{2}}} \\
&\leq \sigma_{\max}
\end{aligned}$$

Pre $y = (1, 0, \cdots, 0)$, $\|\Sigma y\|_2 = \sigma_{\max}$ sa suprémum nadobúda.

□

Vráťme sa teraz k tvrdeniu z druhej kapitoly o zosilnení efektu multikolinearity matice X v matici $X^\top X$. Jednoducho nahliadneme, že opäť pre 'rozumné' normy

$$c(X^\top X) = \|X^\top X\|\|X^{-1}X^{-\top}\| \leq \|X\|^2\|X^{-1}\|^2 = c(X)^2.$$

Z vety je jasné, že $c(X) \geq 1$ a rovnosť nastáva, iba ak sú všetky singulárne hodnoty matice X rovnaké. Pre prípad štvorcovej matice su teda rovnaké aj všetky jej vlastné hodnoty - matica X musí byť nenulový násobok jednotkovej matice (nahliadneme zo spektrálneho rozkladu). Čím horšie sa matica X invertuje, tým je $c(X)$ vyššie. Invertovateľnosť matice ale úzko súvisí s lineárnou závislosťou jej stĺpcov, ktorú v reči štatistiky môžeme zjednodušene chápať ako koreláciu. Preto ak je číslo podmienenosti zásobníka c(X) vysoké, zásobník s veľkou pravdepodobnosťou obsahuje veľa korelovaných premenných, ktoré sú z hľadiska vybudovania stabilného modelu nadbytočné, alebo dokonca až nežiaduce. Kvalitný zásobník teda bude mať toto číslo blízke jednotke, pretože jeho prediktory sú dostatočne dekorelované.

Môže sa zdať, že použiť toto kritérium nemusí byť na mieste. Ak by sme totiž za X dosadili jednotkovú maticu vhodných rozmerov, $c(X)$ bude rovné dokonca jednej, ale predikčná schopnosť nášho regresného modelu značne utrpí z dôvodu pretrénovania a nulového súvisu prediktorov s predikovanou premennou. To však hrozí iba v prípade, že by sa počet neurónov v skrytej vrstve približoval k počtu pozorovaní, čo však obyčajne nikdy neplatí, pretože kým pozorovaní sú rádovo tisícky, počet neurónov by sa mal na-

staviť tak, aby veľmi nepresahoval počet prediktorov, ktoré fenomén reálne vysvetľujú, čo sú rádovo desiatky až stovky.

4.2 RMSE

Keďže hlavnou úlohou našich modelov je v prvom rade presne predikovať, potrebujeme aj kritérium, ktoré je schopné ohodnotiť, ako dobre predpovedáme cieľovú premennú Y. Za týmto účelom sa používa klasická odmocnina strednej kvadratickej chyby (root mean square error). Označme predikované hodnoty ako $\hat{Y}$. Potom $RMSE$ vypočítame ako:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y - \hat{Y})^2}{n}}.$$

5 Vzťah čísla podmienenosti a topológie W

5.1 Dátové sety

5.1.1 Spotreba plynu

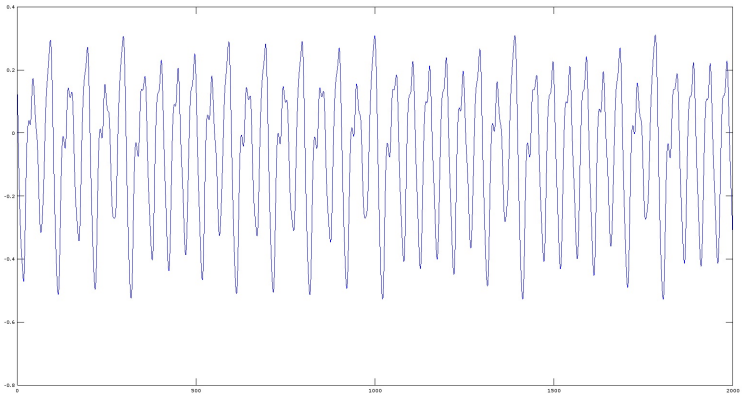
Dáta obsahujú jednu cieľovú vysvetľovanú premennú - spotrebu plynu v okolí jedného z veľkomiest. Okrem toho obsahujú štyri prediktory - oblačnosť, rýchlosť vetra, teplotu a daždivosť. Všetky dáta boli vycentrované na nulovú strednú hodnotu a znormované na jednotkovú varianciu. Jedná sa o jedno vykurovacie obdobie merané každú hodinu, čo dáva dokopy približne 3500 dát. Okrem týchto štyroch prediktorov používame ešte piaty vstup, ktorým je samotná spotreba plynu, ale posunutá o 24 hodín dozadu.

5.1.2 Mackey-Glass

Tieto dáta sú umelo vyrobené a chronicky používané pri testovaní kvality väčšiny modelov určených pre predikciu časových radov. Vychádzajú z diferenciálnej rovnice

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1+x(t-\tau)^{10}} - bx(t).$$

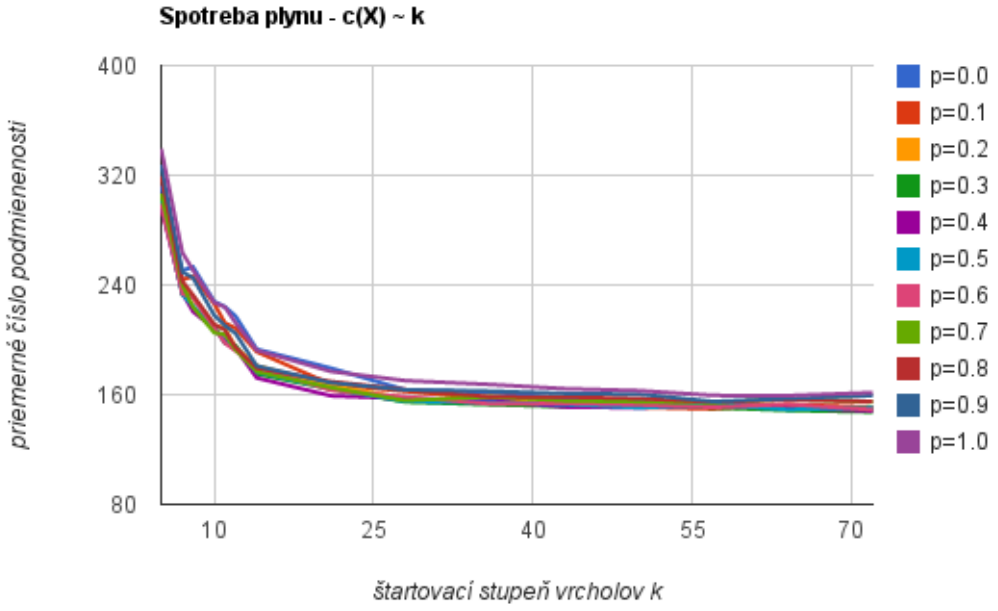
V našich pokusoch sme použili konfiguráciu $a = 0.2, b = 0.1$ a $\tau = 17$. Vygenerovali sme 2000 dát s časovým krokom 0.1 a počiatočnou podmienkou $x_0 = 1.2$, časť časového radu môžeme vidieť na obrázku. Ako vstup do modelu používame ten istý signál, ale posunutý o jednu časovú jednotku dozadu.



Obr. 12: Mackey-Glass časový rad

5.2 SW siete

V tejto časti si popíšeme, ako prebiehali simulácie, ktorých cieľom bolo zistiť, ako je ovplyvnená kvalita zásobníka hodnotou parametrov vstupujúcich do tvorby SW siete. Pre zafixované hodnoty k - počiatočný stupeň vrcholov a p - pravdepodobnosť preskupenia hrany sme najprv vygenerovali maticu W^{in}, pričom jej prvky boli generované z rovnomerného rozdelenia $(-1/2, 1/2)$. Následne sme 100 krát vygenerovali topológiu siete prostredníctvom algoritmu uvedenom v tretej kapitole, čo nám dalo symetrickú maticu W zloženú z núl a jednotiek a konektivitou $k/(n-1)$. Táto matica mala 144 riadkov a stĺpcov, čo odpovedá 144 neurónom v skrytej vrstve. Po vytvorení topológie sme každú jednotkovú hranu náhodne ovážili číslom z rovnomerného rozdelenia $(-1/2, 1/2)$ (uvedomme si, že tento krok vlastne graf dezorientuje). Maticu W sme znormovali jej najväčšou vlastnou hodnotou. Vytvorili sme časové rady v jednotlivých neurónoch a podľa vzťahov a pravidiel v prvej kapitole skonštruovali zásobník X. Nakoniec sme pre každú zo sto realizácií vypočítali číslo podmienenosti matice X. Ako výsledok sme potom pre každú dvojicu vstupných parametrov ukladali priemernú hodnotu týchto čísel podmienenosti a ich štandardnú ochýlku. Na grafoch a v tabuľkách v prílohe môžeme vidieť, ako to dopadlo. Tu uvádzame iba jeden z nich na ilustráciu.



Obr. 13

Jasne vidíme, že číslo $c(X)$ klesá so zvyšujúcim sa číslom k. Pre dáta spotreby plynu sa zrejme navyše dá povedať, že pravá SW topológia (t.j. p=0.4 až 0.6) dopadla lepšie ako pravidelná a úplne náhodná štruktúra, ale tieto dohady sa pre Mackey-Glass nepotvrdili. Čo však povedať môžeme, je, že používať hodnotu $p = 0.5$ bude dávať rozumné výsledky v oboch prípadoch. V tomto bode teda vznikla otázka, ako súvisí $c(X)$ od k, ktorej sme sa v ďalších pokusoch venovali. Dôležité je si uvedomiť, že číslo k v skutočnosti označuje omnoho viac vecí, pokiaľ nepoužívame SW topológiu. Môže tak ísť o priemerný počet stupňov vrcholov v grafe, tak isto môže k zohrávať úlohu ako priamo úmerná konektivita $k/(n-1)$ a tiež môže mať tento parameter význam iba pre SW siete presne v takom zmysle, ako sme ho použili. Prvé dve možnosti nám našťastie splynú do jednej, pretože priemerný stupeň v grafe sa počas algoritmu zachová a ide teda o číslo k/n, ktorého súvis s $c(X)$ bude samozrejme rovnaký ako súvis čísla $k/(n-1)$. Najprv sa ale pozrime na to, akú veľkú úlohu v tomto procese zohráva znáhodnenie hrán a či odpoveď na našu otázku zásadne neovplyvňuje.

5.2.1 Dôležitosť náhodných váh

V nasledujúcich pokusoch sme sa obmedzili na $p = 0.5$. Pozreli sme sa na to, čo sa so zásobníkom stane, ak vynecháme krok ováženia hrán náhodným číslom. Taktiež sme sa pozreli na to, čo sa stane, ak túto náhodnosť síce ponecháme, ale obmedzíme na alternatívne rozdelenie $Alt(\{-1, 1\}, p = 0.5)$. Nakoniec sme k tomuto znáhodneniu pridali ešte podmienku o zachovaní pôvodnej symetrie matice, čo v praxi znamená, že hrany medzi neurónmi ostali neorientované. Z týchto výsledkov sa dá vyčítať mnoho rozličných závislostí.

Tabuľka 1: Priemerné číslo podmienenosti pre rôzne typy náhodného ováženia hrán - plyn

priemerný stupeň	sw + unif	sw	sw + alt	sw + alt + sym
k = 7	233.0719809	581.1101581	194.2536367	380.3495218
k = 14	176.4849049	763.8436375	166.9402604	369.9252434
k = 21	164.280637	900.0557554	154.90388	365.8951035
k =28	155.1075809	1029.250913	153.7643751	369.1729269
k =36	154.3171215	1140.213713	151.5280022	371.9607388
k =43	152.9638883	1222.252907	153.1984454	363.376613
k =50	150.5165668	1306.501481	148.4532121	361.7024933
k =57	152.7201613	1358.015224	147.9258251	367.2832034
k =64	149.1791785	1454.363194	149.3611377	363.4712511
k =72	150.4302692	1545.386612	148.7972807	368.3217394

1. Snáď najviac zjavnou je, že ak maticu W necháme symetrickú s jednotkami a nulami (stĺpec sw), model sa stane veľmi nestabilným. Najpravdepodobnejším vysvetlením pre tento jav je častý zdroj nestability popísaný v druhej kapitole, ktorý súvisí s 'pumpovaním' kladných hodnôt do sigmoidy, ktoré postupne produkujú všetky čísla veľmi blízke jednotke. Nejde tak o nič, čo by sme nečakali.

2. Druhým zaujímavým pozorovaním je, že ako zdroj náhodnosti môžeme používať iba samotné alternatívne rozdelenie (stĺpec sw + alt) a číslo podmienenosti bude k minimu konvergovať so zvyšujúcim sa k dokonca rýchlejšie, ako keby sme použili rovnomerné rozdelenie (stĺpec sw + unif).

3. Výsledok v poslednom stĺpci navráva, že rozlíšiť hrany medzi dvoma neurónmi má tiež veľký význam. Dôvodom, prečo je to tak, sa budeme bližšie venovať neskôr.

4. A nakoniec, najdôležitejším výsledkom je pre nás potvrdenie domienky, že ak je náhodnosť nastavená rozumne - t.j. kladné aj záporné znamienka vo W, naozaj môžeme pozorovať pokles čísla podmienenosti so stúpajúcim parametrom k. V ďalších simuláciách sa pozrieme, či sa na k máme pozerať ako parameter SW siete, alebo na priemerný stupeň neurónov v skrytej vrstve.

Tabuľka 2: Priemerné číslo podmienenosti pre rôzne typy náhodného ováženia hrán - MG

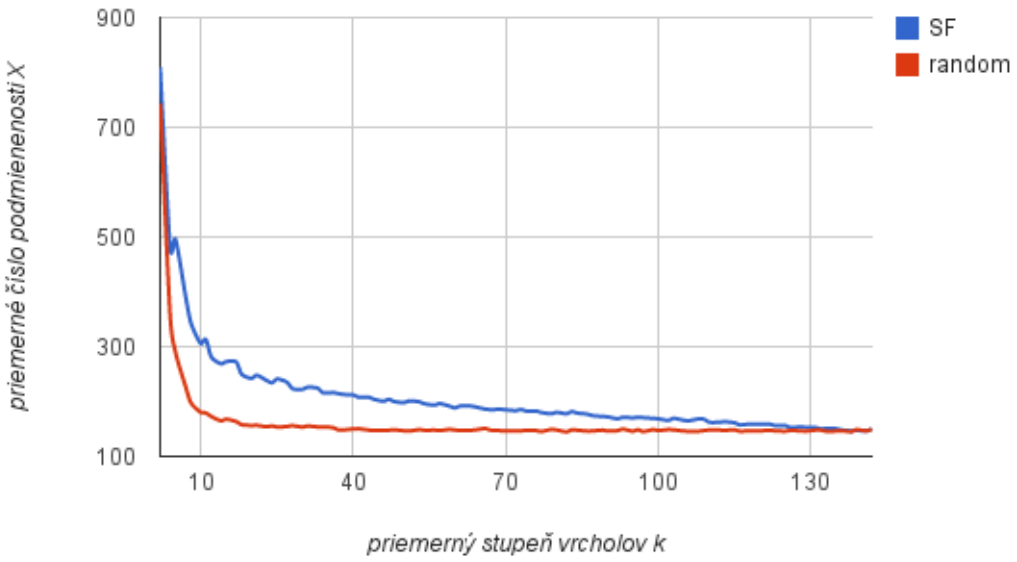
priemerný stupeň	sw+unif	sw	sw+alt	sw+alt+sym
k = 7	1329413.62	90403022.59	1056410.155	12619538.76
k =14	651929.9391	283078984.2	561308.4682	12979679.02
k =21	459925.4886	569501205.2	446051.8919	13663602.82
k =28	371319.5065	1000815116	374593.5295	13351841.42
k =36	358841.323	1662208761	368124.581	13752794.08
k =43	362671.7678	2471634086	371982.1041	13705879.94
k =50	329669.3883	3389580424	335145.3033	13320628.7
k =57	347733.0923	4351740544	326510.0946	14201177.18
k =64	332763.6276	5744782193	341155.298	14253529.86
k =72	356113.2664	7595460207	316744.7705	14052300.84

5.3 SF a náhodná sieť

V nasledujúcich simuláciách sme namiesto SW siete použili najprv SF, kde sme menili iba parameter k, ktorý v tomto prípade znamená priemerný stupeň vrcholov v grafe. Hrany sme ovážili opäť pomocou rovnomerného rozdelenia. Ako druhú topológiu sme zvolili úplne náhodnú maticu, ktorej konektivita bola zvolená tak, aby sme dosiahli odpovedajúci priemerný stupeň vrcholov. V nej sme hrany opäť ovážili pomocou rovnakého rozdelenia. Rozdiel medzi týmito dvoma topológiami je predovšetkým v rozdelení stupňov vrcholov. Označme ako náhodnú premennú stupeň náhodne vybraného vrchola. Kým v prvom prípade bude jej rozdelenie približne exponcenciálne (s takým cieľom je sieť konštruovaná), v druhom pôjde približne o binomické (ak by sme uvažovali možnosť viacnásobných hrán, bolo by presne binomické, takto sa so zvyšujúcim počtom hrán bude pravdepodobnosť postupne meniť na istotu pre stupeň $n-1$) - obe však s rovnakou strednou hodnotou (priemerom k).

Vidíme, že pozorujeme ten istý fenomén ako pri SW sieťach - číslo podmienenosti klesá so stúpajúcim k. Teraz je už zrejmé, že nepôjde o charakteristiku topológie, ale na k by sme sa mali pozerať naozaj iba ako na priemer - nie ako konštrukčný parameter. Hypotéza teda znie, že číslo podmienenosti matice klesá so stúpajúcim priemerným

stupňom vrcholov v grafe W. V ďalšej kapitole sa pozrieme na možné príčiny, ktoré všetky nami odhalené javy vysvetľujú.



Obr. 14

6 Diskusia

6.1 Dezorientácia hrán vo W

V predchádzajúcej kapitole sme si všimli, že ak pri danej topológii ovážime hrany smerujúce medzi dvoma neurónmi rovnako, dostaneme značne horšie výsledky, ako keby sme ponechali znamienko náhodné. Pozrime sa teda na zjednodušený model v nasledujúcom tvare, kde a a b reprezentujú dva neuróny:

$$a(i+1) = 0.5(a(i) + S_1 b(i)) + \varepsilon_a$$

$$b(i+1) = 0.5(S_2 a(i) + b(i)) + \varepsilon_b,$$

pričom chyby ε pochádzajú zo štandardizovaného normálneho rozdelenia a S_i nadobúda hodnoty 1 a -1.

Tabuľka 3: korelácie vektorov a a b

S1/S2	+	-
+	0.98685	-0.023701
-	0.015928	-0.97549

Podľa očakávaní sme dostali veľmi korelované časové rady pri rovnakých znamienkach a pekne dekorelované rady pri rozličných. Je jasné, že táto závislosť sa potom duplikuje po celej sieti a pri zlom nastavení spôsobuje veľké číslo podmienenosti matice X.

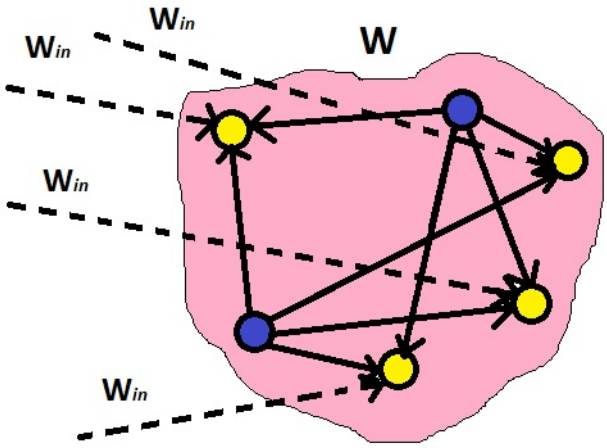
6.2 Malý stupeň vrcholov ako zdroj multikolinearity

V tejto časti sa pokúsime vysvetliť, prečo sú vrcholy s malým stupňom výrazným zdrojom multikolinearity. Dôvody sú dva. Stupňom vrchola budeme myslieť počet prichádzajúcich hrán (pre potreby tejto časti bude postačujúce vychádzajúce hrany neuvažovať).

6.2.1 Málo neurónov v zhluku

Malý stupeň vrcholu - neurónu - v podstate hovorí, že neurón dostáva informáciu iba zo vstupov a malej hŕstky iných neruónov, ktoré sú s ním spojené. Ak však aj jeho

susedné neuróny dostávajú rovnako málo informácie, potom sa ľahko môže stať, že časové rady v týchto susedoch budú korelovať, keďže zo zvyšných hrán nedostávajú dostatočnú veľkosť variability. Tento problém sa obzvlášť objaví, ak sa v našej sieti nachádzajú zhluky, ako sme mohli vidieť pri SW sieťach pre malé hodnoty k. Ak teda v sieti existujú zhluky, ale neuróny v nich majú malý stupeň, potom signály v zhluku nebudú dostatočne dekorelované.



Obr. 15: žlté neuróny dostávajú v zhluku veľmi podobnú informáciu

6.2.2 Vplyv hrán z W^{in}

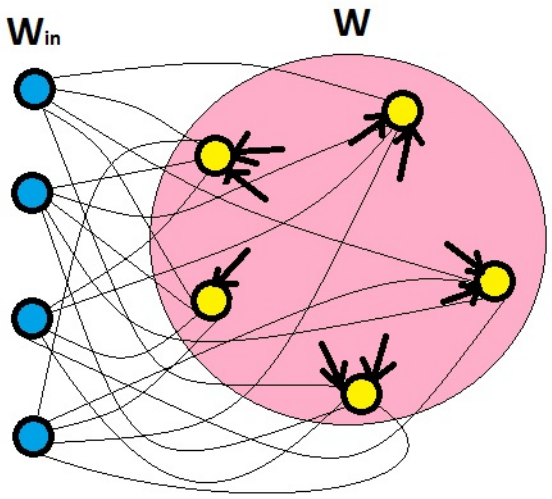
Ďalším možným problémom je zrejme multikolinearita medzi množinou vrcholov s rovnako malým stupňom. Uvažujme na chvíľu iba neuróny so stupňom 2. Označme túto množinu ako W_2. Predpokladajme model so štyrmi vstupmi. Potom daný neurón z našej množiny W_2 dostáva informáciu z dvoch iných neurónov v skrytej vrstve a štyroch neurónov zo vstupnej vrstvy. Ak signál z neurónov v skrytej vrstve (vrátane neurónu samotného) stotožníme s náhodným šumom, potom aktualizácia stavu v takomto neuróne vyzerá ako

$$x(n+1) = f(c_{x1}u_1 + c_{x2}u_2 + c_{x3}u_3 + c_{x4}u_4 + \varepsilon_x).$$

Pre iný neurón z W_2 sa zmení iba ‘rozdelenie‘ ε a váhy c_{ij}. Teraz si stačí uvedomiť, že ak je neurónov vo W_2 veľa, ich časové rady musia byť ‘lineárne závislé‘ (v zmysle lineárnej aktivačnej funkcie f^{out} a vynechania chyby ε). Báza takéhoto priestoru má totiž iba

4 vektory. Navyše do W_2 môžeme uvažovať aj analogicky definovanú W_1, ktorá je v tomto zmysle jeho podmnožinou. Neuróny v týchto množinách potom výrazne prispejú k multikolinearite.

Takýto druh multikolinearity sa potom detekuje omnoho ťažšie, pretože lineárna závisloť nie je to isté ako korelácia, aj keď spolu súvisia. Totiž vektor, ktorý sa dá napísať ako lineárna kombinácia zvyšných, stále nemusí mať vysokú koreláciu s jednotlivými vektormi v báze. Pre jednoduchosť si predstavme vektory $(1,0,0),(0,1,0),(0,0,1)$ a $(1,1,1)$. Potom posledný z nich je zjavne lineárne závislý od ostatných, lebo tvoria bázu, ale s jednotlivými bázovými vektormi zviera väčší ako 30 stupňový uhol, čiže jeho korelácia s nimi nebude podozrivo vysoká. To znamená, že obyčajnou korelačnou analýzou pomocou kovariančnej matice tento zdroj multikolinearity neodhalíme.

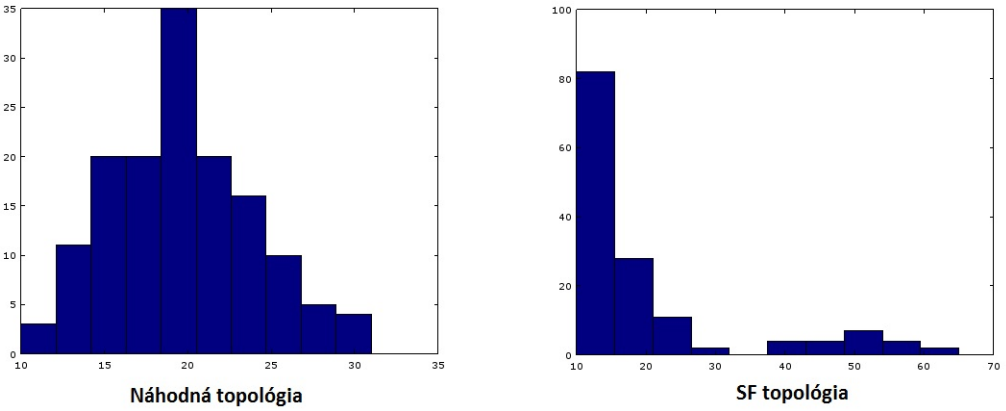


Obr. 16: žlté neuróny dostávajú zo vstupu korelovanú informáciu

Je dôležité uvedomiť si, že toto platí naozaj iba ak i vo W_i je malé. Totiž ak by tomu tak nebolo, 'chyba' ε je v skutočnosti nosnou časťou signálu v neuróne x a nemôžeme ju len tak zanedbať. Zrejme hranicu pre 'malosť' i vieme nastaviť iba v závislosti od počtu prediktorov vstupujúcich do modelu, lebo práve tento počet ovplyvňuje veľkosť bázy, ktorú sme uvažovali. Do hry taktiež vstupuje aj magnitúda náhodných váh hrán.

Tieto domienky krásne podporuje posledná simulácia minulej kapitoly, kde sme porovnávali, ako klesá číslo podmienenosti matice X s rastúcim priemerným stupňom vrcholov. Môžeme si všimnúť, že síce obidve závislosti boli klesajúce, ale pre SF sieť

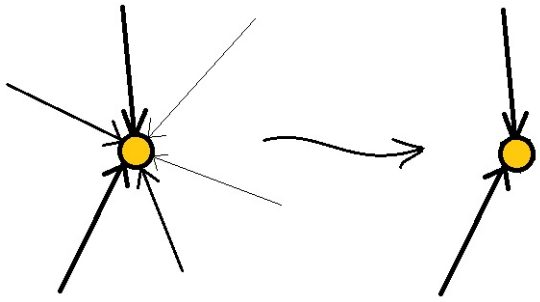
bol tento trend značne pomalší. Je to preto, lebo ako už bolo spomenuté, kým rozdelenie stupňov vrcholov pre náhodnú topológiu je približne binomické (t.j. symetrické), rozdelenie stupňov pre SF sieť je približne exponcenciálne (aspoň pre nižšie hodnoty k). To v praxi znamená, že aj keď priemer stupňov vrcholov ostáva pre obe topológie rovnaký, pre SF bude počet vrcholov s malým stupňov oveľa vyšší, a teda hranica k, kde už všetky vrcholy nadobúdajú dostatočný stupeň, je oveľa vyššie ako pre náhodnú topológiu. Ďalším javom, ktorý toto tvrdenie podporuje, je zistená rýchlejšia konvergencia čísla podmienenosti X k minimu, ak rovnomerné rozdelenie nahradíme alternatívnym.



Obr. 17: histogramy stupňov vrcholov v sieti so 144 neurónmi v skrytej vrstve pre $k = 20$

6.3 Prečo alternatívne rozdelenie predčí rovnomerné

Odpoveď na túto otázku sa zrejme nachádza už v predošlej časti. Ak totiž použijeme na náhodné ováženie hrán rovnomerné rozdelenie, niektoré neuróny v sieti nadobudnú efektívny stupeň nižší, ako je reálny počet vstupujúcich hrán do neurónu. Efektívnym stupňom sa tu myslí, do akej miery tvoria signály prichádzajúcich hrán výsledný signál v neuróne. Totiž, ak je realizácia váhy blízka nule, spojenie medzi neurónmi síce existuje, ale je takmer zanedbateľné. Takže naše topológie obsahovali pre rovnaké k v skutočnosti ‘viac‘ neurónov s malým stupňom pre rovnomerné rozdelenie ako pre alternatívne, kde je váha každej rany v absolútnej hodnote rovnaká, a teda nezanedbateľná.



Obr. 18: reálny versus efektívny stupeň neurónu

6.4 Zhrnutie výsledkov

1. Hrany skrytej vrstvy by mali obsahovať kladné aj záporné znamienka. Ak ponecháme iba jeden druh, vďaka 'pumpovaniu' hodnôt do chvostov sigmoidálnej aktivačnej funkcie sa nám kvalita zásobníka zhorší.

2. Ak medzi dvoma neurónmi vedú hrany oboma smermi, je lepšie ich ovážiť opačným znamienkom. Tento zásah potom pomôže časové rady v neurónoch navzájom dekorelovať.

3. Bez ohľadu na to, aký typ topológie používame, mali by sme si dať pozor, aby neobsahovala príliš veľa neurónov s malým prichádzajúcim stupňom, pretože ako sme demonštrovali, tie sú zrejme jedným zo zdrojov multikolinearity.

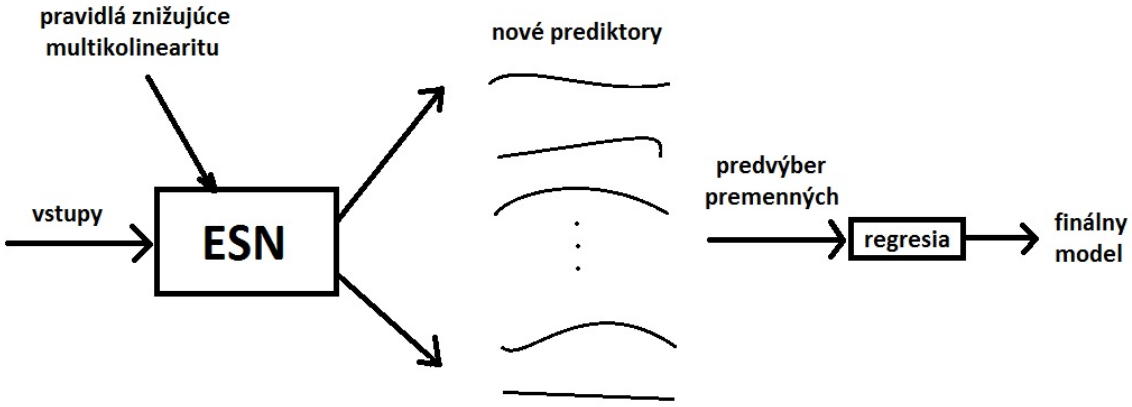
4. Siete pri rovnakom počte hrán dosahujú lepšie výsledky z hľadiska čísla podmienenosti zásobníka X, ak namiesto ováženia hrán realizáciami zo spojitej hustoty pravdepodobnosti použijeme jednoduché alternatívne rozdelenie. Deje sa tak, pretože týmto spôsobom vieme umelo zvýšiť počet relevantných vchádzajúcich hrán do neurónov.

6.5 Prediktívne vlastnosti modelu a ďalšie obzory

Všetky naše úvahy sme robili z hľadiska čísla podmienenosti matice X. Ako sme však už uviedli, rozhodujúcou črtou pre naše modely by mala byť ich schopnosť predikovať. Pri pokusoch s predikciami sme ale narazili na problém, kedy síce topológie, ktoré sa držali daných štyroch vlastností, dosahovali lepšie výsledky (z hľadiska RMSE), ale

tento rozdiel nebol natoľko signifikantný, aby sa dal jasne vypichnúť. Dôvodov môže byť viacero. Ako sme už uviedli, dobré číslo podmienenosti nemusí hneď zaručovať lepšie výsledky pri predikciách. Zároveň sú takto vytvorené modely značne 'pretrénované', pretože dané fenomény sa zaručene dajú vysvetliť omnoho nižším počtom relevantných prediktorov, ako sme používali pri skúmaní kvality zásobníka. V klasickej lineárnej regresii totiž do trénovania vstupujú všetky pomocou ESN vygenerované prediktory.

Napriek tomu má zmysel mať kvalitný zásobník, pretože prístup ESN sa dá kombinovať s metódami, ako je LASSO, hrebeňová regresia a dopredný výber premenných, ktoré vhodným spôsobom penalizujú nepotrebné prediktory, prípadne okliesnia množinu premenných iba na kvalitnejšiu podmnožinu. Táto cesta sa dá brať ako samostatný výskum, ktorý použitím spomínaných techník odhalí, do akej miery číslo podmienenosti zásobníka spôsobí väčšiu alebo menšiu kvalitu predikcií.



Obr. 19: schéma použitia našich výsledkov s predvýberom

Záver

Cieľom tejto diplomovej práce bolo oboznámiť sa s problematikou echo sietí, grafovými štruktúrami používanými pri ich tvorbe a ich vlastnosťami. Taktiež bolo cieľom podľa možností aj vyskúmať zaujímavé zákonitosti vplyvu topológie siete na kvalitu vyprodukovaných regresorov. Tento výskum však veľmi rýchlo vyprodukoval množstvo ďalších zaujímavých výsledkov a následných otázok, ktoré autora možno nakoniec trošku odklonili od pôvodnej cesty a viac ako topológii samotnej sa venoval bádaniu, prečo sa číslo podmienenosti zásobníka znižovalo s rastúcou konektivitou v grafe bez ohľadu na to, akú topológiu použil. Pri ceste za odpoveďou na túto otázku sme postupne prišli na ďalší rad zaujímavých pozorovaní - pri spojitom ovažovaní hrán bol tento pokles pomalší, rozličné znamienka vo váhach majú obrovský význam a symetria váh medzi neurónmi škodí kvalite zásobníka.

Všetky tieto odsimulované výsledky a cesty hľadania odpovedí na hypotézy nakoniec poviedli k tej istej možnej príčine - vrcholy s malým stupňom zrejme spôsobujú multikolinearitu, ktorá je hlavným nepriateľom všetkých algoritmov strojového učenia. Všetky naše úvahy sú síce iba nepriameho charakteru, ale sú silno podporené sumiláciami.

Toto však zďaleka nemusí byť koniec bádania, pretože vytvoriť dobrý model s prediktívnymi vlastnosťami porovnateľnými so špičkovými algoritmami je úloha, ktorá je oveľa ťažším orieškom. Preto má zmysel nami dosiahnuté výsledky skúmať ďalej a pozrieť sa na aplikácie ESN v spojení s inými algoritmami, ktoré by ho mohli posunúť takýmto spôsobom ešte ďalej. ESN sa v súčasnosti teší v technickom svete veľkej obľube a podlieha neprestajnému výskumu. Preto veríme, že by dosiahnuté výsledky mohli v budúcnosti niekomu pomôcť pri skvalitnení jeho práce.

Zoznam použitej literatúry

[1] Clauset, A., Cosma, R. S., Newman, M. E. J.: *Power-law distributions in empirical data*, SIAM Review, vol. 51, no. 4, pp. 661-703 (7.6. 2007)

[2] Cohen, R., Havlin, S.: *Scale-free networks are ultrasmall*, Phys. Rev. Lett. 90 (5): 058701 (2003)

[3] Dolinský J., Konishi, S.: *Echo State Networks with Non-monotonous Activation Functions*, The 2010 Japanese Joint Statistical Meeting

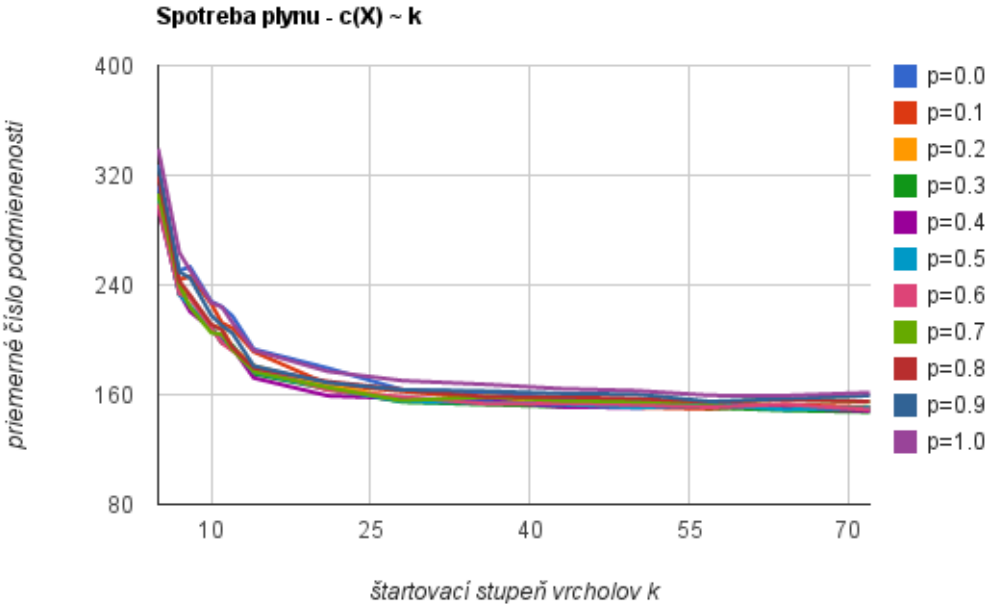
[4] Jaeger H.: *The 'echo state' approach to analysing and training recurrent neural networks*, GMD Report 148 (28. 12. 2001), German National Research Center for Information Technology

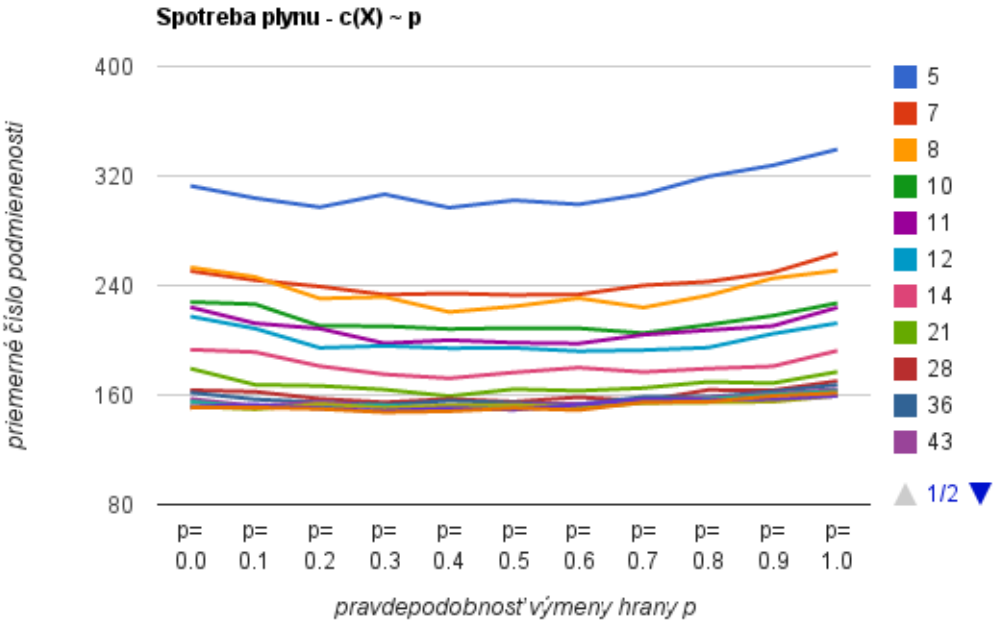
[5] Lukoševičius M.: *A Practical Guide to Applying Echo State Networks*, Neural Networks: Tricks of the Trade, Reloaded (2012), Springer

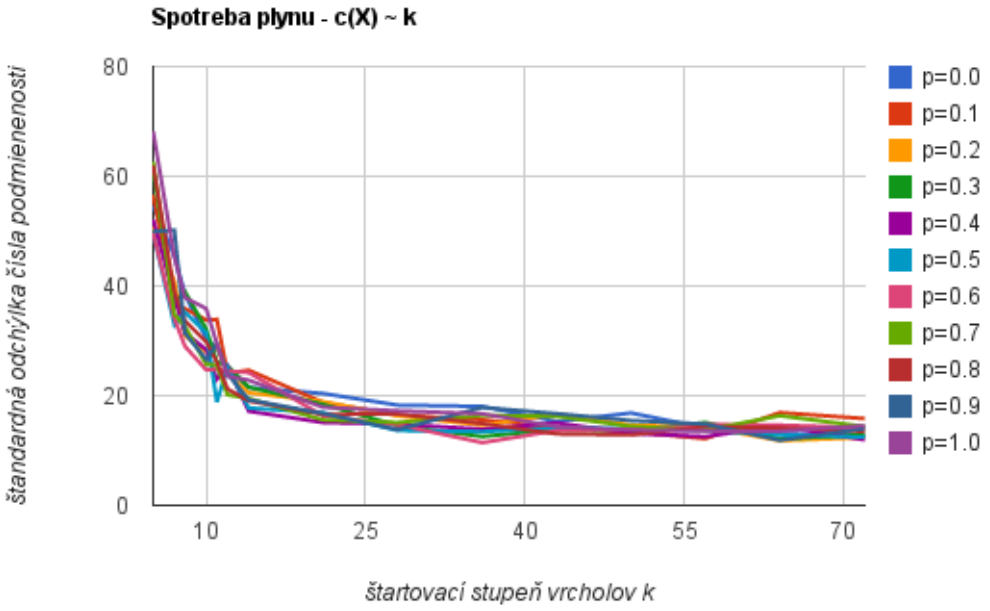
[6] Watts, D. J., Strogatz, S. H.: *Collective dynamics of 'small-world' networks*, Nature Journal Vol 393 (4. 6. 1998), Macmillan Publishers Ltd

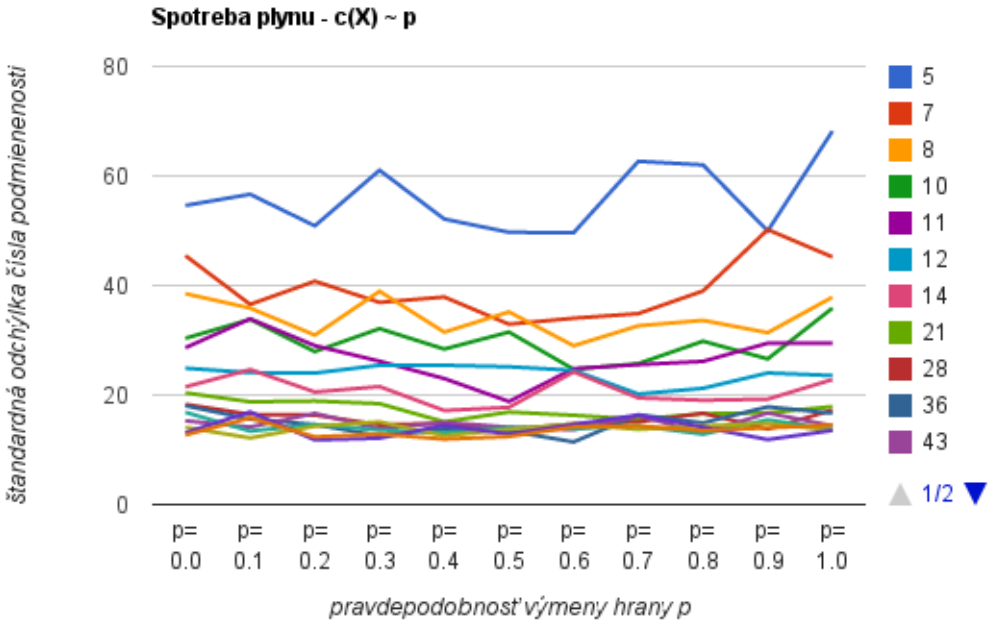
[7] Zhang Y., Deng, Z.: *Collective Behavior of a Small-World Recurrent Neural System With Scale-Free Distribution*, IEEE Transactions of neural networks Vol 18, NO. 5 (September 2007)

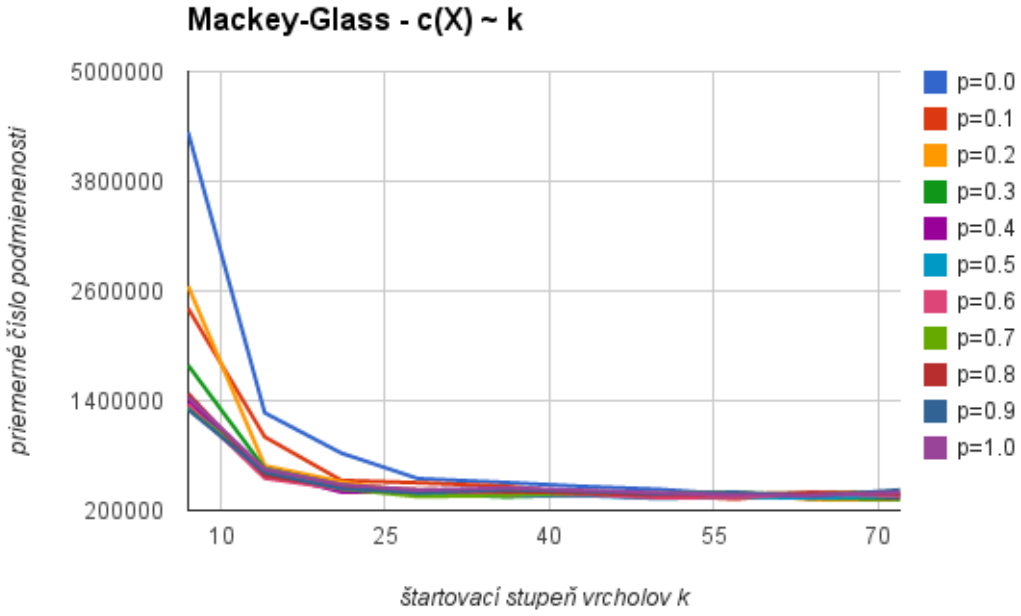
Grafy a tabuľky

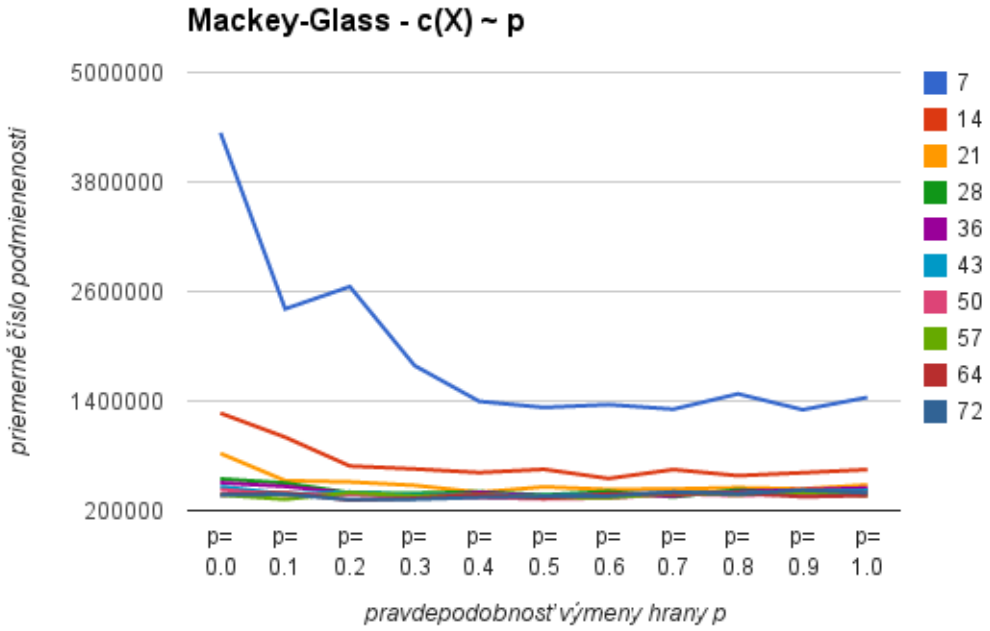


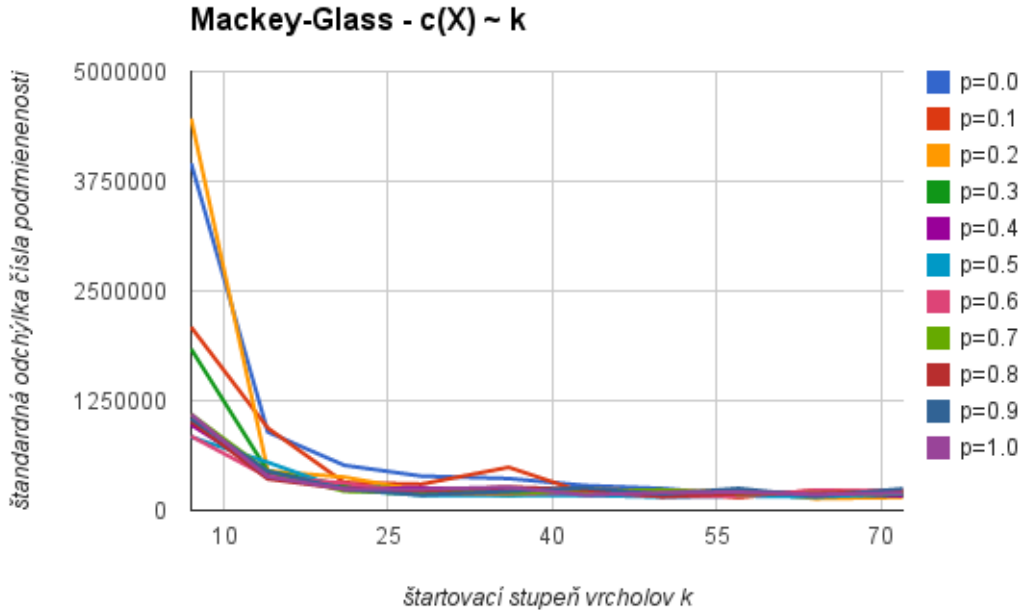


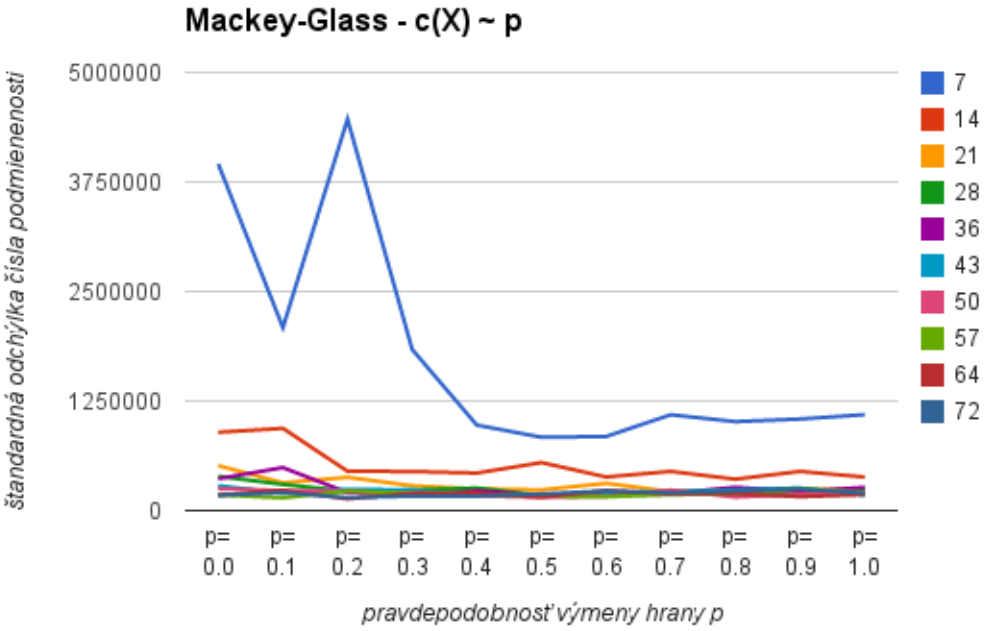

Spotreba plynu - c(X) ~ k
štandardná odchýlka čísla podmienenosti
80
60
40
20
0
10
25
40
55
70
štartovací stupeň vrcholov k
p=0.0
p=0.1
p=0.2
p=0.3
p=0.4
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
p=1.0


Spotreba plynu - c(X) ~ p
štandardná odchýlka čísla podmienenosti
80
60
40
20
0
p= 0.0
p= 0.1
p= 0.2
p= 0.3
p= 0.4
p= 0.5
p= 0.6
p= 0.7
p= 0.8
p= 0.9
p= 1.0
pravdepodobnosť výmeny hrany p
5
7
8
10
11
12
14
21
28
36
43
1/2


Mackey-Glass - c(X) ~ k
priemerné číslo podmienenosti
5000000
3800000
2600000
1400000
200000
10
25
40
55
70
štartovací stupeň vrcholov k
p=0.0
p=0.1
p=0.2
p=0.3
p=0.4
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
p=1.0


Mackey-Glass - c(X) ~ p
priemerné číslo podmienenosti
5000000
3800000
2600000
1400000
200000
p= 0.0
p= 0.1
p= 0.2
p= 0.3
p= 0.4
p= 0.5
p= 0.6
p= 0.7
p= 0.8
p= 0.9
p= 1.0
pravdepodobnosť výmeny hrany p
7
14
21
28
36
43
50
57
64
72


Mackey-Glass - c(X) ~ k
štandardná odchýlka čísla podmienenosti
5000000
3750000
2500000
1250000
0
10
25
40
55
70
štartovací stupeň vrcholov k
p=0.0
p=0.1
p=0.2
p=0.3
p=0.4
p=0.5
p=0.6
p=0.7
p=0.8
p=0.9
p=1.0


Mackey-Glass - c(X) ~ p
štandardná odchýlka čísla podmienenosti
5000000
3750000
2500000
1250000
0
p= 0.0
p= 0.1
p= 0.2
p= 0.3
p= 0.4
p= 0.5
p= 0.6
p= 0.7
p= 0.8
p= 0.9
p= 1.0
pravdepodobnosť výmeny hrany p
7
14
21
28
36
43
50
57
64
72

Tabuľka 4: priemerné čísla podmienenosti pre rôzne parametre SW - spotreba plynu

k	p=0.0	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=1.0
5	313	304	297	307	297	302	299	307	320	328	339
7	251	244	239	233	234	233	233	240	243	249	264
8	253	246	230	232	221	225	231	224	233	245	251
10	228	226	210	210	208	209	209	205	211	218	227
11	224	212	209	198	200	198	198	204	207	210	224
12	217	209	194	196	194	194	192	193	195	205	212
14	193	191	181	175	172	176	180	177	179	181	192
21	179	168	167	164	159	164	163	165	170	169	177
28	164	162	157	155	157	155	158	156	164	163	170
36	162	157	155	153	156	154	153	158	158	163	167
43	157	152	155	151	151	153	154	156	158	161	164
50	155	151	152	152	150	151	153	155	157	160	163
57	152	150	154	151	153	153	151	154	155	155	159
64	152	153	150	149	150	149	153	157	156	157	159
72	151	151	150	147	148	150	149	155	155	159	162

Tabuľka 5: priemerné čísla podmienenosti pre rôzne parametre SW - MG (dáta v tisíckach)

k	p=0.0	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	p=0.9	p=1.0
7	4339	2409	2653	1788	1396	1329	1361	1310	1478	1305	1440
14	1267	1004	688	656	616	652	551	649	582	615	649
21	826	528	514	476	401	460	431	437	451	438	481
28	547	502	394	390	405	371	410	351	428	392	422
36	505	465	390	348	398	359	370	364	407	428	448
43	463	380	403	369	368	363	375	394	390	410	419
50	431	369	375	349	341	330	337	384	366	384	392
57	372	324	398	359	372	348	342	388	374	399	366
64	374	397	315	336	376	333	384	383	393	352	366
72	371	379	319	323	345	356	360	404	381	425	394

Tabuľka 6: priemerné čísla podmienenosti a ich štandardné odchýlky pre SF a náhodnú topológiu NT - spotreba plynu

k	SF	NT	SF - std	NT - std
2	809.1972041	743.2335544	168.0272664	192.1975594
7	392.4262273	227.620512	79.13382188	23.10661397
11	313.0558049	179.013548	54.34058532	16.67264796
15	272.6063693	167.8300214	38.63049758	13.19610387
20	242.3510054	156.3778456	28.31853259	12.36875936
24	234.3051468	155.4103422	27.39456523	12.56725127
28	223.5638248	155.964578	22.46629795	14.55582439
33	224.2867201	153.6821759	26.26138126	14.3548363
37	214.628948	148.5244671	24.37446215	12.93119835
41	207.9157154	150.4178368	19.63742014	14.05332737
46	200.5629167	147.5801398	21.07631216	15.79027802
50	198.0673009	146.6296691	18.9350586	12.78000077
54	195.9939464	147.6659267	20.52687653	12.38926775
59	191.3958638	149.1598566	20.09483705	13.59549567
63	191.9434374	147.4526866	17.1754926	13.04468932
67	185.3374456	147.8611437	17.52006715	12.76214064
72	183.491504	146.6368396	20.23883152	13.31838213
76	182.6120156	146.9847682	18.17683263	14.80636118
80	180.3613375	148.118368	16.75597369	11.72669994
84	178.9800717	147.7947308	17.23171571	13.5784941
89	172.7402185	147.8765983	15.72350453	13.12041487
93	171.088927	149.7527127	13.81873498	13.71163578
97	170.6607896	145.1445142	19.09359013	11.31081779
102	165.5105514	148.8392262	12.41851285	13.09598797
106	164.9974035	145.9853231	16.20354952	11.6744952
110	162.232379	148.0748186	13.01801467	12.85703516
115	161.1764391	148.4679496	12.23801039	12.38942138
119	158.7391836	146.4976112	14.36281469	12.99705191

Zdrojové kódy pre tvorbu topológií

Uvedené kódy boli písané pre program Octave, ale mali by fungovať aj pre Matlab.

B.0.1 Small World

```
function G = Nsw(N,degree,p)
 G = sparse(zeros(N,N));
 Ninitial=floor(floor(degree+.5)/2);
 for i = 1:N
  for link = 1:Ninitial
    G(i,mod(i+link-1,N)+1)=1;
  end;
 end;
 k=floor(degree+.5)/2;
 if(floor(k)<k)
  for i =1:2:N
    G(i,mod(i+floor(k),N)+1)=1;
  end;
 end;

 for i=1:N
  for j=i+1:N
    if(G(i,j)==1 && rand()<p)
      G(i,j)=0;
      a=floor(rand()*N)+1;
        while(G(a,i)==1 || G(i,a)==1 || i==a)
          a=floor(rand()*N)+1;
        end;
      if(i<a)
        G(i,a)=1;
      else
        G(a,i)=1;
      end;
    end;
  end;
```

```
end;

if degree<floor(degree+.5)
 Nremoved=0;
 Ntoremove=floor((floor(degree+.5)-degree)*N);
 while(Ntoremove>Nremoved)
   for i=1:N
     for j=i+1:N
       if(Ntoremove>Nremoved && G(i,j)==1 && rand()<Ntoremove/(Npervertice*N))
         G(i,j)=0;
         Nremoved=Nremoved+2;
       end;
     end;
   end;
 end;
else
 Nadded=0;
 Ntoadd=floor((degree-floor(degree+.5))*N);
   while(Ntoadd>Nadded)
     for i=1:N
       for j=i+1:N
         if(Ntoadd>Nadded && G(i,j)==0 && rand()<Ntoadd/(Npervertice*N))
           G(i,j)=1;
           Nadded=Nadded+2;
         end;
       end;
     end;
   end;
end;
G = G+G';
```

B.0.2 Scale free

```
function G = Nsf(N,d)
 G = zeros(N,N);
 added = zeros(N,1);
 for i = 1:(d+1)
  for j = (i+1):(d+1)
    G(i,j) = 1;
    G(j,i) = 1;
  end;
  added(i) = 1;
 end;

 for i = (d+2):N
  for l = 1:(d/2)
    p = (G*added).*added.*(ones(N,1)-G(:,i));
    p = p/(ones(1,N)*p);
    s = rand;
    m = 1;
    while (s>p(m))
      s = s-p(m);
      m = m+1;
    end;
    G(m,i) = 1;
    G(i,m) = 1;
  end;
  added(i) = 1;
 end;
 G = sparse(G);
```

B.0.3 Tvorba obrázkov

```
function drawNet(G)

format compact
format long e
a = linspace(0,2*pi,length(G)+1);
z = zeros(length(G)+1,2);
x = cos(a);
y = sin(a);
z(1:length(G)+1,1) = x(1:length(G)+1);
z(1:length(G)+1,2) = y(1:length(G)+1);
figure; gplot(G,z,'.-');
set(gcf, 'Color', [1 1 1]);
axis('equal');
xlim([-1.1 1.1]);
ylim([-1.1 1.1]);
axis off;
```

Techniky kombinovateľné s ESN

C.0.4 Ortogonálny dopredný výber

Ako sme uviedli v závere poslednej kapitoly, častým problémov lineárnej regresie v modelovaní býva veľké množstvo premenných, ktoré spôsobujú pretrénovanie. Tento problém sa nás obzvlášť dotýka, pretože ESN môže teoreticky vygenerovať ľubovoľné množstvo prediktorov v závislosti od počtu neurónov v skrytej vrstve. Úloha výberu podmnožiny vysvetľovacích premenných preto má zmysel a je dobré poznať techniky vyvinuté na jej riešenie.

Ortogonálny dopredný výber stojí na jednoduchej myšlienke, že vidieť jednotlivé príspevky prediktorov v modeli je oveľa jednoduchšie pri ich ortogonálom dizajne. Ak aj naše prediktory nie sú navzájom kolmé, vždy ich môžeme prostredníctvom Gram-Schmidtovho ortogonalizačného procesu vhodne sprojektovať tak, aby boli. Využime QR rozklad matice $X_{n\times m}$, avšak s tým rozdielom, že vektory v báze nebudeme normovať:

$$Y = X\beta + \varepsilon$$

$$Y = QR\beta + \varepsilon$$

$$Y = Qg + \varepsilon,$$

pričom Q je ortogonálna a $g = R\beta$. Obe strany rovnice umocníme:

$$Y^\top Y = g^\top Q^\top Q g + \varepsilon^\top \varepsilon.$$

Keďže $Q^\top Q$ je diagonálna, dostávame

$$Y^\top Y = g_1^2 q_1^\top q_1 + g_2^2 q_2^\top q_2 + \cdots + g_m^2 q_m^\top q_m + \varepsilon^\top \varepsilon$$

$$1 = \sum_{i=1}^{m} err_i + \frac{\varepsilon^\top \varepsilon}{Y^\top Y}$$

$$err_i = \frac{g_i^2 q_i^\top q_i}{Y^\top Y}.$$

Ak zlomok v err_i rozšírime o počet pozorovaní n, tak za predpokladu nulového priemeru Y dostávame, že err_i udáva podiel vysvetlenej variancie daným prediktorom. Toto sa dá v praxi využiť tak, že výber premenných stopneme, ak je súčet vysvetlenej variancie

dostatočne veľký, t.j. prekračuje nejakú vopred stanovenú hodnotu ξ. Celý algoritmus teda vyzerá ako Gram-Schmidtova ortogonalizácia s tromi úpravami:

- vektory nenormalizujeme
- vektory do matice Q nevyberáme v pôvodnom poradí, ale vždy na základe najväčšieho err_i
- výber stopneme, ak kumulovaná suma err_i vybraných vektorov prekročí vopred stanovenú hodnotu ξ.

Najväčšou slabinou tohoto algoritmu je samozrejme nemožnosť stanoviť ξ generický pre všetky typy dát. Preto sa v praxi používajú na výber najlepšieho vektora v danom kroku rôzne informačné kritéria alebo iné druhy metrík. Jednou z nich je PRESS kritérium.

C.0.5 PRESS

Skratka PRESS pochádza z anglického 'predicted residual sum of squares', čo môžeme preložiť ako predikovaný súčet štvorcov odchýliek. Toto kritérium v sebe spája myšlienky cross-validácie a tradičného súčtu štvorcov odchýliek, ktorý poznáme z lineárnej regresie. Definujeme ho ako

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2,$$

kde $\hat{y}_{i,-i}$ značí hodnotu predikcie modelu pre i-tu hodnotu, ak bol model vybudovaný pomocou všetkých hodnôt okrem i-tych. Ide teda o klasickú leave-one-out cross-validáciu použitú počas trénovania modelu. Obrovskou výhodou tohoto kritéria je, že sa na rozdiel od kumulovaného err_i nespráva monotónne s pribúdajúcim počtom prediktorov v modeli. Vychádza to z podstaty cross-validácie, ktorá by mala zabrániť pretrénovaniu modelu. Ak ju takto šikovne implementujeme už v rámci samotného budovania modelu (na rozdiel od tradičného porovnávania s testovacou vzorkou odloženou bokom), máme k dispozícii silné kritérium, kedy výber stopnúť. Ak totiž PRESS krivka (meniaca sa hodnota PRESS kritéria s pribúdajúcim počtom prediktorov v modeli) začne stúpať, je to indikátor toho, že v modeli je zrejme už dostatočne veľa vysvetlenej variancie. Nevýhodou tohoto kritéria je samozrejme výpočtová zložitosť, ale existujú

programátorské skratky vychádzajúce z algebraických úprav, ktoré ju vedia značne skresať. Tie sú však už nad rámec tejto diplomovej práce.