

Optimal Brain Damage

(Dokumentácia k 2. zadaniu)

Bc. Peter Brtáň

Študijný program: Softvérové inžinierstvo

Ročník: 2.

Predmet: Neurónové siete

Ak.rok:2008/2009

Obsah

1. Úvod	1
2. Definícia problému.....	1
3. Dôvody výberu problému	1
4. Spôsob riešenia problému – Optimal Brain Damage.....	1
5. Algoritmické riešenie problému.....	3
6. Experimenty	4
7. Záver.....	5

1. Úvod

Predkladaný dokument je dokumentáciou k 2.zadaniu – individuálnemu projektu z predmetu Neurónové siete. V dokumente je stručne definované samotne zadanie projektu, dôvody prečo je zaujímavé venovať sa danému problému a spôsob akým je možné problém vyriešiť. V záverečných častiach dokumentu je uvedené algoritmičné riešenie načrtnutého problému, štatistiky získane použitím vytvorenej aplikácie a závery vyplývajúce z týchto štatistík. V samom závere sú načrtnuté aj možné vylepšenia vytvoreného algoritmu v budúcnosti.

2. Definícia problému

Cieľom projektu je vytvoriť doprednú neurónovú sieť so spätným šírením chýb, ktorá bude schopná optimalizovať svoju štruktúru. Problém možno teda definovať aj ako nájdenie spôsobu, akým by bolo možné odstrániť „nadbytočne“ spojenia v sieti a týmto spôsobom teda upraviť neurónovú sieť tak, aby jej „kvalita predikcie“ bola stále akceptovateľná a súčasne jej štruktúra čo najoptimálnejšia.

3. Dôvody výberu problému

Mnoho úspešných aplikácií neurónových sietí aplikovaných na reálne problémy z praxe bolo dosiahnutých použitím vysoko štruktúrovaných sietí veľkých rozmerov. Ako sa aplikácie stavajú zložitejšími, siete rastú ešte viac a stavajú sa ešte štruktúrovanejšími. Z týchto dôvodov sú potrebné nástroje na porovnanie rôznych architektúr neurónových sietí a nástroje na optimalizáciu štruktúry neurónových sietí.

Problémy z veľkým množstvom spojení v neurónových sieťach sú vysoké výpočtové nároky a dlhý čas potrebný na natrénovanie sietí a problém zo všeobecnosťou neurónových sietí. Všeobecnosť neurónových sietí je ich schopnosť poskytnúť správny výsledok pre príklady, ktoré nie sú obsiahnuté v tréningovej množine. Je všeobecne známe, že najlepšie výsledky neurónové siete dávajú pri ich optimálnej štruktúre. Ak je počet spojení v sieti príliš vysoký dochádza k tzv. „overfitting“ problému, kedy sieť od určitého počtu spojení poskytuje stále horšie výsledky zovšeobecnovania. Naopak ak je počet spojení v sieti príliš nízky nemá dostatočnú silu na to, aby poskytovala správne výsledky pre príklady, ktoré sa nenachádzajú v tréningovej množine. Preto je potrebné nájsť správny pomer medzi chybovosťou siete a jej zložitosťou.

4. Spôsob riešenia problému – Optimal Brain Damage

Záverom predchádzajúcej kapitoly je potreba nájsť správny pomer medzi chybovosťou siete (priemerná chyba po natrénovaní siete) a jej zložitosťou. Jednou z techník ako to dosiahnuť je minimalizácia nákladovej funkcie zloženej z dvoch parametrov: štandardná chybovosť siete a určitá miera zložitosti siete. Medzi navrhnuté mieri zložitosti patria napr. Vapnikova-Chervonenkisova dimenzionalita [1] a dĺžka popisu [2]. Pre účely tohto projektu bola zvolená

metóda počtu nenulových spojení [3]. Vo väčšine prípadov najmä z oblasti štatistickej inferencie existuje určitá a priority alebo heuristická informácia, ktorá určí poradie spojení, ktoré by mali byť odstránene z neurónovej siete. Príkladom môže byť sústava polynomických rovníc, kedy heuristika môže požadovať najprv odstrániť členy z najvyšším radom. Vo väčšine prípadov však vôbec nie je známe v akom poradí by mali byť spojenia zo siete odstránene.

Jednoducho stratégiou je odstránenie spojení s najmenšou „dôležitosťou“. Po odstránení je potrebné danú sieť znovu natrénovať a tento proces môže byť opakovaný pokiaľ neurónová sieť stále ponúka uspokojivé výsledky. Pri tomto prístupe však nastáva problém ako určiť „dôležitosť“ jednotlivých spojení. Podľa [3] je možno „dôležitosť“ spojení definovať ako nárast chybovej funkcie siete po ich odstránení. Na výpočet dôležitosti spojení je použitá druhá derivácia chybovej funkcie E podľa jednotlivých spojení v sieti. Chybová funkcia je použitá najmä z dôvodu jej ústrednej úlohy pri dopredných neurónových sieťach zo spätným šírením chýb.

Priame naplnenie tohto postupu (postupne odstránenie jednotlivých spojení a znovu vyhodnotenie chybovej funkcie) je však neuskutočniteľne kvôli výpočtovej náročnosti tohto postupu. Na druhej strane je však možné zostaviť lokálny model chybovej funkcie a analyticky predpokladať vplyv odstránenia jednotlivých spojení na chybovú funkciu. Vplyv narušenia vektora spojení (δU) na chybovú funkciu je možné definovať nasledovne:

$$\delta E = \sum_i g_i \delta u_i + \frac{1}{2} \sum_i h_{ii} \delta u_i^2 + \frac{1}{2} \sum_{i \neq j} h_{ij} \delta u_i \delta u_j + O(\|\delta U\|^3) \quad (1)$$

kde $\delta u(i)$ sú komponenty δU , $g(i)$ sú komponenty gradientu G funkcie E podľa U a $h(i)(j)$ sú elementy Hessianvej matice H funkcie E podľa U:

$$g_i = \frac{\partial E}{\partial u_i} \quad \text{and} \quad h_{ij} = \frac{\partial^2 E}{\partial u_i \partial u_j} \quad (2)$$

Cieľom je ako už bolo napísane nájsť množinu takých parametrov, ktorých odstránenie zo siete spôsobí najnižší nárast chybovej funkcie E. Tento problém je taktiež nemožne prakticky vyriešiť, keďže matica H má obrovské rozmery (pre sieť z 2600 parametrami má 65 000 000 prvkov) a je veľmi náročne ju vyrátať. Z týchto dôvodov muselo byť predstavených niekoľko zjednodušujúcich predpokladov [3]. Diagonálny predpoklad predpokladá, že δE spôsobená odstránením niekoľkých spojení je rovný sume δE –čiek spôsobených odstránením každého spojenia individuálne. Týmto predpokladom môže byť odstránený tretí člen pravej strany rovnice (1). Extrémny predpoklad predpokladá, že spojenia sú vymazané až po natrénovaní siete, kedy k vektoru spojení je priradená minimálna hodnota E a teda prvý člen pravej strany rovnice (1) môže byť odstránený. Nakoniec kvadratický predpoklad predpokladá, že nákladová funkcia je takmer kvadratická, takže aj posledný člen rovnice môže byť odstránený.

Rovnica (1) sa tak redukuje na:

$$\delta E = \frac{1}{2} \sum_i h_{ii} \delta u_i^2$$

Po nájdení rovnice na výpočet „dôležitosti“ jednotlivých spojení v sieti je ešte potrebné nájsť spôsob na výpočet druhej derivácie pri výpočte prvkov Hessianovej matice. Nasledovne vzorce sú uvedené pre jeden príklad z tréningovej množiny. Tieto vzorce je tak ešte potrebné „spriemerovať“ pre celú tréningovú množinu. Stav siete je možné vyrátať pomocou štandardných vzorcov:

$$x_i = f(a_i) \quad \text{and} \quad a_i = \sum_j w_{ij} x_j \quad (4)$$

kde $x(i)$ je stav neurónu i , $a(i)$ je jeho „net“ hodnota (váhovaná suma spojení), f je aktivačná funkcia a $w(i,j)$ je spojenie idúce z neurónu j do neurónu i . Diagonálne prvky matice H sú dane nasledovným vzorcom:

$$h_{kk} = \sum_{(i,j) \in V_k} \frac{\partial^2 E}{\partial w_{ij}^2} \quad (5)$$

Predchádzajúci vzorec môže byť rozšírený pomocou vzorcov z rovnice (4) nasledovne:

$$\frac{\partial^2 E}{\partial w_{ij}^2} = \frac{\partial^2 E}{\partial a_i^2} x_j^2 \quad (6)$$

Druhé derivácie sú potom šírene naspäť medzi vrstvami:

$$\frac{\partial^2 E}{\partial a_i^2} = f'(a_i)^2 \sum_l w_{li}^2 \frac{\partial^2 E}{\partial a_l^2} + f''(a_i) \frac{\partial E}{\partial x_i} \quad (7)$$

Na výpočet druhej derivácie pre neuróny vo výstupnej vrstve sa používa nasledujúci vzorec:

$$\frac{\partial^2 E}{\partial a_i^2} = 2f'(a_i)^2 - 2(d_i - x_i)f''(a_i) \quad (8)$$

5. Algoritmické riešenie problému

V skratke možno algoritmické riešenie načrtnutého problému opísať nasledovne:

1. Voľba určitej architektúry neurónovej siete
2. Natrénovanie siete, až na želanú úroveň predikcie

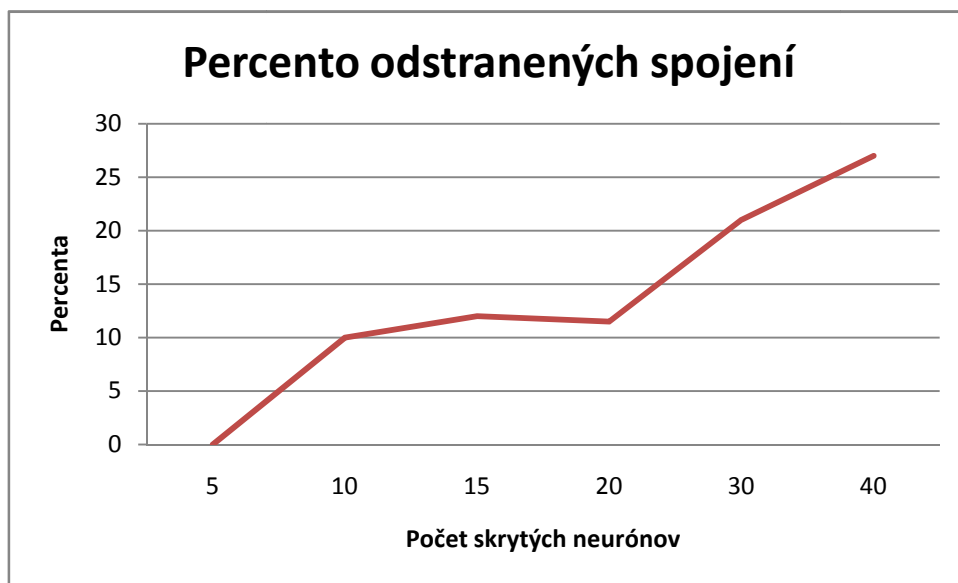
3. Vyrátanie druhej derivácie $h(i,j)$ pre každé spojenie
4. Vyrátanie dôležitosti pre každé spojenie: $s(i,j) = h(i,j)*w(i,j)^2/2$
5. Zoradenie parametrov podľa dôležitosti a odstránenie najmenej dôležitých parametrov
6. Pokračovanie v kroku 2.

Pod odstránením spojenia sa rozumie jeho nastavenie na nulu a zmrazenie – nepoužívanie pri ďalšom používaní a tréovaní siete.

6. Experimenty

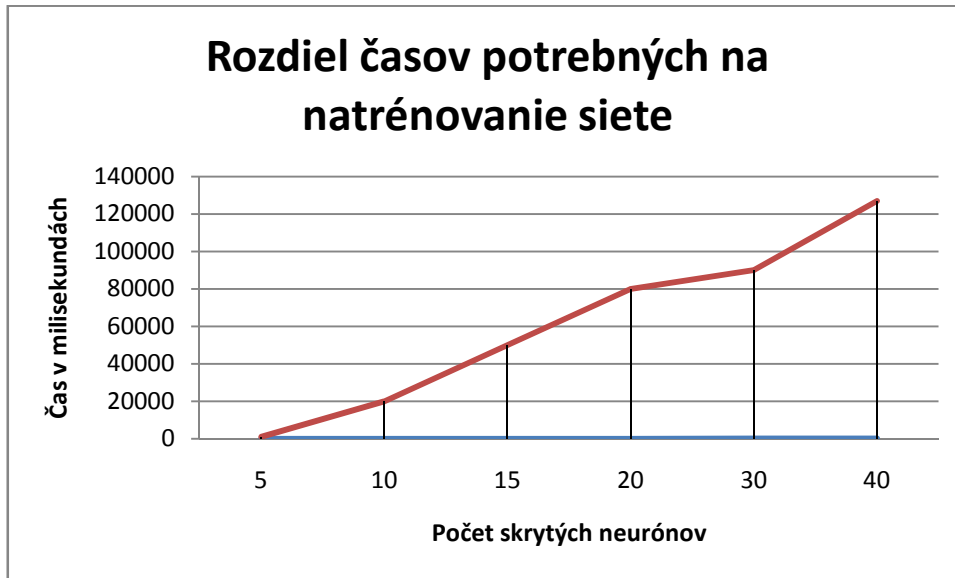
V tejto kapitole uvádzame dva experimenty vykonane na vytvorenej aplikácii. Všetky experimenty boli vykonane na príklade binárnej parity známeho aj z 1.zadania.

Prvým experimentom je percento odstránených spojení (počet odstránených spojení/počet všetkých spojení) zo siete zo zvyšujúcim sa počtom skrytých neurónov v sieti. V každom optimalizačnom cykle sa odstránia 3 najmenej dôležité spojenia. Ako je možné vidieť percento zredukovaných spojení vzrastá s veľkosťou optimalizovanej siete až na 27%. Správnosť použitia algoritmu Optimal Brain Damage je tak zjavná. Aj po zredukovaní siete o viac ako štvrtinu bola kvalita predikcie rovnaká ako na začiatku.



Na druhom grafe je možné vidieť rozdiel časov v milisekundách potrebných na natrénovanie siete na začiatku pri úplnej sieti a na konci po optimalizácii. Výsledky tohto experimentu nezodpovedajú pôvodnému predpokladu, že po optimalizácii sa bude neurónová sieť trénovať rýchlejšie. Naopak pri každom pokuse sa sieť po optimalizácii trénovala pomalšie. Po odstránení spojení zo siete sa ich hodnota nastavuje na nulu a neskôr v kóde sa porovnávajú na rovnosť z nulou. Toto je jedna z možností spomalenia učenia sa. Rovnaký prístup však bol zvolený aj autormi algoritmu Optimal Brain Damage, ktorý taktiež pripúšťajú možnosť spomalenia učenia sa siete. Celkový rozdiel týchto časov však bol v najhoršom

prípade len 2 minúty, čo taktiež nie je znepokojujúci údaj. Spomalenie tréningovania sa siete o 2 minúty je akceptovateľný údaj ak sa vezme prínos pri redukcii siete až o 25%.



7. Záver

Na základe výsledkov predstretých v predchádzajúcej kapitole je možné konštatovať, že metóda Optimal Brain Damage je vhodnou alternatívou pre optimalizovanie neurónových sietí. Ako možné vylepšenie tohto problému do budúcnosti vidíme úpravu algoritmu tak, aby nebol zřejmý rozdiel časov potrebných na natrénovanie siete na začiatku pri úplnej architektúre a po optimalizácii siete. Možným vylepšením aplikácie je zase nájdenie testovacieho príkladu z väčším počtom tréningových dát a celkovo väčšou zložitosťou siete.

8. Použitá literatúra

[1] Vapnik, V. N. a Chervonenkis, A. Y. (1971). On the Uniform Convergence of Relative Frequencies of Events To Their Probabilities. Th. Prob. And its Applications, 17(2):264-280.

[2] Rissanen, J. (1989). Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore.

[3] Le Cun, Y., Denker J. S. a Solla S. A. Optimal Brain Damage. AT&T Bell Laboratories.