Neuronske mreže za detekciju Spam-a

Jelena Križman
Davor Menon

24.05.2007.

1 O bazi podataka...

U sklopu kolegija Računarski praktikum 3 radi se projekt iz neuronskih mreža. Za taj projekt je potrebna baza podataka na kojoj se trenira, unakrsno propituje i testira mreža. Bazu podataka se biralo iz arhive baza podataka sa University of California Irwine [1].

Baza, sa pripadajućim dodatnim datotekama, se može skinuti na adresi [3]. Darovatelj je George Forman. Baza je stvorena 1999. godine.

Baza se koristila kao interni izvještaj u kompaniji Hewlett-Packardi sa tendencijom javnog objavljivanja. Korištena je za određivanje koji mail je spam, a koji nije spam. U praksi se pokazalo da je grešla u klasifikaciji bila približno 7%. Važno je razumjti da je pogrešna klasifikacija poruke kao spam vrlo nepoželjna. Pokazalo se da ukoliko se u treniranju ne tolerira pogrešna detekcija spama, 20% - 25% spama prolazi kroz sito.

U bazi je 4601 unosa - od kojih je 39.4% spam. Svaki unos ima 57 neprekidnih atributa i jednu nominalnu oznaku klase. Prije izrade neuronskih mreža se oznaku klase "rastavilo" u dvije klase (posebno "spam", a posebno "no spam"). Svaki atribut je numerička reprezentacija broja određenih riječi u poruci. Baza nema nepotpunih vrijednosti.

Ulazni atributi su sljedeći:

word_freq_make	word_freq_address	word_freq_all
word_freq_3d	word_freq_our	word_freq_over
word_freq_remove	word_freq_internet	word_freq_order
word_freq_mail	word_freq_receive	word_freq_will
word_freq_people	word_freq_report	word_freq_addresses
word_freq_free	word_freq_business	word_freq_email
word_freq_you	word_freq_credit	word_freq_your
word_freq_font	word_freq_000	word_freq_money
word_freq_hp	word_freq_hpl	word_freq_george
word_freq_650	word_freq_lab	word_freq_labs
word_freq_telnet	word_freq_857	word_freq_data
word_freq_415	word_freq_85	word_freq_technology
word_freq_1999	word_freq_parts	word_freq_pm
word_freq_direct	word_freq_cs	word_freq_meeting
word_freq_original	word_freq_project	word_freq_re

```
word_freq_edu       word_freq_table         word_freq_conference
char_freq_;         char_freq_(             char_freq_[
char_freq_!         char_freq_\$            char_freq_#
capital_run_length_average                  capital_run_length_longest
capital_run_length_total
```

Izlazni atributi su:

```
Spam            NoSpam
```

2 Što je spam?

Oko definicije spama lome se još uvijek mnoga koplja, pogotovo otkad je cijeli slučaj dobio svoju pravnu dimenziju. Spomenimo neke definicije spama (u kontekstu elektroničke pošte):

- Spam je preplavljivanje Interneta (flooding) velikim brojem kopija iste poruke, s ciljem da se poruka dostavi onima koji bi, uz mogućnost izbora, odabrali ne primiti je.
- Spam predstavlja svaku email poruku koju korisnik dobije, a koja nema direktne ili indirektne veze s njim. Pod direktnim vezama podrazumijevaju se osobe, tvrtke i razni pružatelji usluga sa kojima korisnik komunicira, a pod indirektnim vezama mislimo na osobe, tvrtke i pružatelje usluga koje se pozivaju na direktne veze (tj. slučajevi gdje direktne veze mogu potvrditi da su proslijedile email adresu).
- Spam je svaka poruka za koju ne postoji razlog da se pojavi u mailboxu

Osim email spama, druga glavna kategorija je usenet spam, odnosno slanje iste poruke na dvadesetak različitih news grupa.

Više o spamu možete pogledati na [2].

Rb	Broj skrivenih slojeva (hidden layers)	broj skrivenih neurona (# Hidden PE)	max. broj epoha za učenje (Epoch)	Prijenosna funkcija u skrivenom i izlaznom sloju (Transfer)	Pravilo učenja (Learning Rule)	Momentum	TRAIN rezultati	TEST rezultati	
							Najniža greška (MSE)	Stopa klasifikacije (za svaku klasu posebno)	Prosječna stopa klasifikacije
1	1	2	1000	SingmoidAxon	DeltaBarDelta	-	0,04416128	81,6156 90,01782	85,8167
2	1	2	2000	SingmoidAxon	DeltaBarDelta	-	0,043646444	81,05 90,02	85,53
3	1	2	3000	SingmoidAxon	DeltaBarDelta	-	0,043415561	81,06 90,02	85,54
4	1	4	1000	SingmoidAxon	DeltaBarDelta	-	0,043244697	82,172699 89,48306	85,8278
5	1	4	2000	SingmoidAxon	DeltaBarDelta	-	0,0437046	81,6156006 90,0178223	85,817
6	1	4	1500	SingmoidAxon	DeltaBarDelta	-	0,0433878	81,0584946 90,0178223	85,5381
7	1	4	3000	SingmoidAxon	DeltaBarDelta	-	0,0434071	81,6156006 90,0178223	85,8167
8	1	8	1000	SingmoidAxon	DeltaBarDelta	-	0,0440787	80,7799454 90,0178223	85,3988
9	1	8	3000	SingmoidAxon	DeltaBarDelta	-	0,0418306	79,665741 91,265594	85,4657
10	1	30	2000	SingmoidAxon	DeltaBarDelta	-	0,0364234	80,2228394 93,9393921	87,0811
11	1	50	1000	SingmoidAxon	DeltaBarDelta	-	0,04129	81,8941498 90,0178223	85,956
12	1	50	2000	SingmoidAxon	DeltaBarDelta	-	0,017797	87,4651794 95,1871643	91,3262
13	1	50	5000	SingmoidAxon	DeltaBarDelta	-	0,04412	85,5153198 86,6702347	85,092
14	1	10	1000	SingmoidAxon	Momentum	0,7	0,046761	81,3370438 89,6613159	85,4992
15	1	10	1000	SingmoidAxon	Momentum	0,7	0,044985	80,5013962 90,0178223	85,2596
16	1	10	1000	SingmoidAxon	Momentum	0,1	0,058942	88,8579407 77,0053482	82,9316
17	1	10	1000	SingmoidAxon	Momentum	0,8	0,046234	80,7799454 89,8395690	85,3098
18	1	10	1000	SingmoidAxon	Momentum	0,99	0,046503	81,8941498 89,6613159	85,7777
19	1	25	1000	SingmoidAxon	Momentum	0,99	0,046810	81,6156006 88,7700501	85,1928
20	1	50	2000	SingmoidAxon	Momentum	0,99	0,040825	82,4512558 90,9090881	86,6802

Tablica 1: Rezultati testiranja

3 Rezultati testiranja

Zbog ograničenja softwarea broj ulaznih varijabli je smanjen na 47. Svaka varijabla je numerička vrijednost koja reprezentira broj pojavljivanja određene

riječi u poruci.

Pošto je ovo problem klasifikacije postoje dvije izlazne varijable, “Spam” i “NoSpam”.

Provedeno je ispitivanje na dvadeset mreža kojima su mijenjani parametri na takav način da se dobije što bolja točnost mreže. Rezultati su prikazani u Tablici 2.

Mijenjani parametri su:

Maksimalan broj epoha za učenje
: Ovaj parametar je mijenjan sa početnih 1000 pa u koracima po 1000 sve do maksimalnih 5000. U nekim slučajevima je to pozitivno utjecalo na prosječnu stopu klasifikacije, ali ponekad je rezultat bio lošiji.

Broj skrivenih neurona
: Ovaj parametar se pokazao kao značajan faktor uspješnosti naših mreža. Vrijednosti su od početnih 2 do maksimalnih 50. Povećavanjem ovog parametra se dobivao puno bolja prosječna stopa klasifikacije. Iako u nekim slučajevima to nije imalo značajnijeg utjecaja.

Pravilo učenja u skrivenom sloju
: U prvih trinaest mreža se koristilo pravilo “DeltaBarDelta”, a u ostalih sedam “Momentum”. Kod “Momentum” pravila smo mijenjali parametar od 0.1 do 0.9. To nije značajno utjecalo na prosječnu stopu klasifikacije.

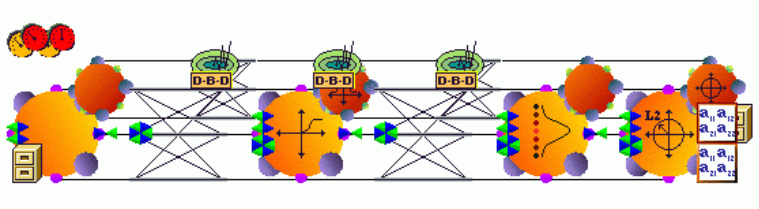
Iz tablice se jasno vidi da je najbolje rezultate postigla mreža pod rednim brojem 12. Ona je postigla prosječnu stopu klasifikacije od $91,3262\%$.

S druge strane, najlošija mreža je pod rednim brojem 16. Ona je postigla prosječnu stopu klasifikacije od $82,9316\%$

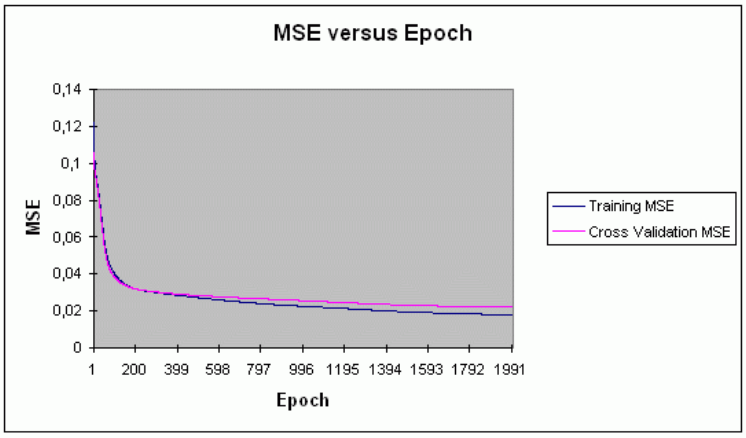
3.1 Mreža s najboljim rezultatima

Parametri u mreži s najboljim rezultatima su:

Redni broj mreže	12
Broj skrivenih slojeva	1
Broj skrivenih neurona	50
Maksimalan broj epoha	2000
Prijenosna funkcija u skrivenom i izlaznom sloju	SigmoidAxon
Pravilo učenja	DeltaBarDelta



Slika 1: Grafička interpretacija najbolje mreže



Best Networks	Training	Cross Validation
Epoch #	2000	1948
Minimum MSE	0,017796528	0,022383649
Final MSE	0,017796528	0,022386592

Slika 2: Grafikon kretanja greške i unakrsne provjere

Iz grafikona kretanja greške (slika 2) i unakrsne validacije vidimo da se nakon 250 iteracija (epoha) grafovi razilaze.

Output / Desired	*Spam*	*NoSpam*
Spam	314	27
NoSpam	45	534

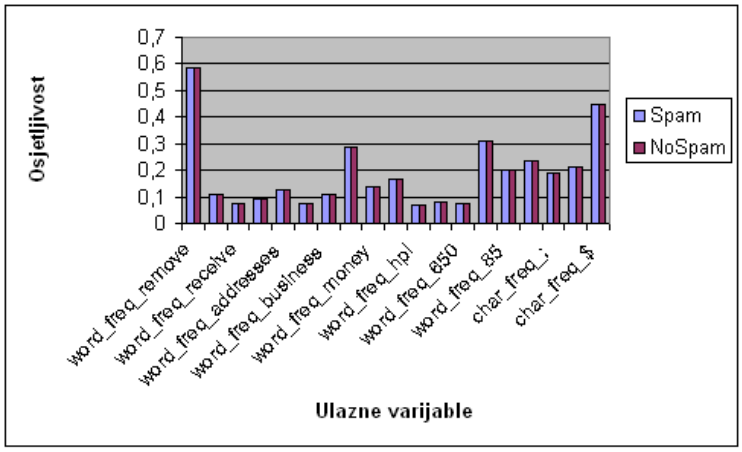
Performance	*Spam*	*NoSpam*
MSE	0.059365402	0.05942993
NMSE	0.249489205	0.249760388
MAE	0.127051203	0.1271893
Min Abs Error	0.000131904	0.000131965
Max Abs Error	1.027436256	1.027436288
r	0.866772747	0.866772749
Percent Correct	87.46517944	95.18716431

Tablica 2: Matrica konfuzije

U gornjem dijelu Tablice 2 je broj slučajeva koji su ispravno i pogrešno klasificirani. U stupcima je prikazan stvarni broj poruka označenih sa "Spam" i "NoSpam". Ukupno je bilo 395 Spam-a u tom uzorku od kojih je njih 45 pogršno svrstano u klasu NoSpam. Broj ispravno klasificiranog Spam-a prikazan je zadnjem retku donjeg dijela tablice (Percent Correct - Spam): 87.47%. U uzorku za testiranje je bilo 561 NoSpam-a od kojih je njih 534 mreža uspjela dobro klasificirati u NoSpam, dok je 27 pogrešno svrstano u Spam. U postotcima to iznosi 95.19%

4 Značajnost ulaznih varijabli

Slika 3 prikazuje kolika je osjetljivost izlaznih varijabli Spam i NoSpam na svaku od ulaznih varijabli. Vrijednosti na y osi grafikona prikazuju kolika je promijena izlaznih varijabli, ako se pojedina ulazna varijabla promijeni za jednu jednicu. Vidljivo je da varijbla `word_freq_remove` naviše utječe na izlazne varijable, varijabla `char_freq_$` malo manje i tako redom. Isti rezultati prikazani su i tablično u tablici 3.



Slika 3: Grafikon osjetljivosti ulaznih varijabli na izlazne u modelu klasifikacije poruka prema Spam-u

Literatura

[1] *http://www.ics.uci.edu/ mlearn/MLSummary.html*, University of California Irwine

[2] *http://www.zpr.fer.hr/predmeti/erg/2004/buco/sto_je_spam.htm*, Zavod za primijenjenu matematiku, Fakultet elektrotehnike i računarstva

[3] M.Hopkins, E.Reeber, G.Forman i J.Suermondt, *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/*, Hewlett-Packard Labs, 1999.

Sensitivity	Spam	NoSpam
word_freq_remove	0,5826714	0,5826714
word_freq_internet	0,10980604	0,10980604
word_freq_receive	0,07506455	0,07506456
word_freq_will	0,08989908	0,08989909
word_freq_addresses	0,12741293	0,12741295
word_freq_free	0,07573172	0,07573171
word_freq_business	0,1089491	0,10894909
word_freq_000	0,28623748	0,28623748
word_freq_money	0,13933311	0,1393331
word_freq_hp	0,1659662	0,1659662
word_freq_hpl	0,06807468	0,06807468
word_freq_george	0,08286571	0,08286571
word_freq_650	0,07205002	0,07205003
word_freq_415	0,30874807	0,30874807
word_freq_85	0,20160386	0,20160387
word_freq_1999	0,23296967	0,23296969
char_freq_;	0,19016807	0,19016804
char_freq_!	0,212321	0,21232101
char_freq_$	0,44857267	0,44857267

Tablica 3: Osjetljivost izlaza na ulazne varijable