Korisnička dokumentacija

Domagoj Bošnjak, Iskra Gašparić

1 Opis problema

SMS poruke i mailovi mogu se podijeliti u dvije skupine - prave poruke (ham) te spam. Svi moderni mail poslužitelji koriste filtere kako bi smanjili širenje lažnih poruka. Cilj ovog projekta je kreirati takav algoritam za klasteriranje poruka. Rad je usredotočen na SMS poruke budući da je takva baza podataka korištena, ali logika iza algoritma je primjenjiva i na mailove (uz modifikacije zbog drugačijeg oblika baze).

2 Predprocesiranje podataka

Kako bismo kreirali algoritme, najprije treba analizirati neka svojstva SMS poruka koja će pomoći u klasifikaciji. U odnosu na e-mailove, ovdje nailazimo na veliku ograničenost. Spam mailovi lako se prepoznaju po gramatičkim greškama, manjku ili višku interpukcijskih znakova, krivom pozicioniranju razmaka, sadržaju naslova i samom sadržaju maila. SMS je kratak tekst, gramatičke greške su prirodne i u ham porukama te se ne poštuju gotovo nikakva pravila. Zato je prva ideja koju smo implementirali da poruke pretvorimo u cjelobrojne nizove (svaki znak je jedinstveno reprezentiran integerom) te poruke promatramo takve - bez uzimanja sadržaja poruke u obzir. Drugi pristup je konkretniji: tokenizacijom poruka (korištenjem Porterovog algoritma) te traženjem ključnih i najčešće korištenih riječi u porukama, one se klasificiraju po kontekstu. Mana je duljina SMS-a (svaka poruka sadrži $100 - 200$ znakova ili manje pa je broj riječi malen). Oba pristupa testirana su u svim algoritmima.

3 K-sredine i algoritam k-najbližih susjeda

Kao što je navedeno ranije, oba algoritma testirana su na dvije vrste značajki. U prvom slučaju pristup je bio direktan: znakovi u poruci pretvoreni su u cijele brojeve. Kao i očekivano takav pristup nije dao posebno dobre rezultate. U drugom slučaju matrica značajki konstruirana je kao TFIDF matrica, ali K-means algoritam nije se pokazao previše pouzdanim niti u tom slučaju. Primjer klasteriranja prvim pristupom:

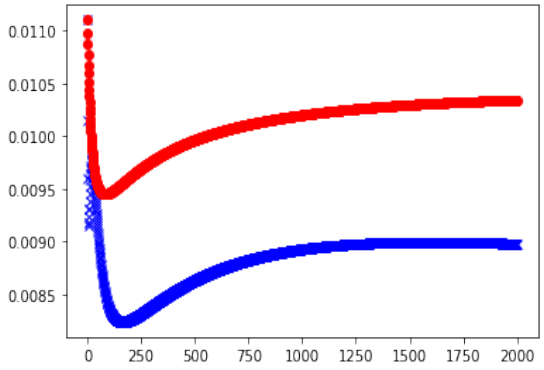
	Prva klasa	Druga klasa
Broj ham poruka	552	901
Broj spam poruka	214	5

S druge strane, k-NN algoritam dao je nešto bolje rezultate, a posebno za slučaj TFIDF matrice. Konkretno, ham SMS poruke separirane su s točnošću preko 99%.

	Prva klasa	Druga klasa
Broj ham poruka	1469	1
Broj spam poruka	79	124

4 Neuronska mreža

Kao funkcija aktivacije odabrana je sigmoidalna funkcija $1/(1+e^x)$. Mreža je jednoslojna, funkcija greške minimizira se gradijentnim spustom. Ulaz u algoritam su poruke te njihova oznaka "ham" ili "spam". Pomoću tih vrijednosti učimo neuronsku mrežu. Ovaj algoritam dao je osjetno bolje rezultate od prethodnih. Čak i prvi pristup problemu (brojčana interpretacija poruka) daje relevantne rezultate. Na uzorku od oko 5000 poruka, od kojih je 70% nasumičnih poruka odabrano kao training set, točnost je oko 90%. Drugi pristup, u skladu s našim pretpostavkama, daje još bolje rezultate. Njegova točnost na istim podatcima doseže i do 94% točnosti! U nastavku je graf na kojem je prikazano ponašanje norme greške (razlika između stvarnih vrijednosti i predviđanja mreže) u svakoj iteraciji. Nakon 1000 iteracija točnost raste dosta sporo i kreće se oko vrijednosti koje su navedene ranije.



Slika 1: Crveno - prvi pristup; plavo - drugi pristup

5 Zaključak

U smislu reprezentacije podataka, bolja rješenja dobivena su kad se za matricu značajki uzela TFIDF matrica, u odnosu na jednostavno pretvaranje slova po-

ruke u cijele brojeve. Dakle, opravdana je pretpostavka da je takav pristup bolji.

K-means algoritam dao je uvjerljivo najlošije rezultate. S obzirom da ne uzima u obzir to da su nam oznake poruka iz training skupa poznate, očekujemo lošije rezultate u odnosu na druga dva algoritma. Klasificiranje neuronskom mrežom i k-NN algoritmom dali su stabilna i relativno točna rješenja.

Literatura

[1] Nilam Nur Amir Sjarif*, Nurulhuda Firdaus Mohd Azmi, Suriayati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, Suriani Mohd Sam:

SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm, The Fifth Information Systems International Conference 2019