

Automatsko prilagođavanje klasifikatora rukom pisanog teksta pojedinačnim korisnicima

Student: Milan Čugurović 1009/2018

Mentor: Mladen Nikolić

Uvod

- Realan problem
- Onlajn/oflajn
- Scenario upotrebe – zahtevi
- Konvolutivne neuronske mreže
- Algoritam K najbližih suseda
- Algoritam klasterovanja K sredina

Osnovne ideje

- Bazni klasifikator: CNN
- ‘Parsirati’ stil pisanja svakog od korisnika
 - Stil pisanja karaktera? (uključena CNN)
- Specifična podela:
 - Trening skup za bazni klasifikator
 - Validacioni skup za bazni klasifikator
 - Skup za primenu:
 - Skup za prilagođavanje
 - Skup za testiranje

Skup podataka

- Problemi
- Nist Special Database 19 (1995)
- ETH Zurich database (2018)
- IAM Handwriting Database?

HANDWRITING SAMPLE FORM

NAME [REDACTED] DATE 8-3-89 CITY Minden City STATE MI ZIP 48956

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below.

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

0123456789 0123456789 0123456789

87 701 3752 80759 960041

87 701 3752 80759 960041

158 4586 32123 832656 82

158 4586 32123 832656 82

7481 80539 419219 67 904

7481 80539 419219 67 904

61738 729658 75 390 5716

61738 729658 75 390 5716

109334 40 625 4234 46002

109334 40 625 4234 46002

g y x l a k p d b t s i r u m w f q j e n h o c v

g y x l a k p d b t s i r u m w f q j e n h o c v

Z X S B N G E C M Y W Q T K F L U O H P I R V D J A

Z X S B N G E C M Y W Q T K F L U O H P I R V D J A

Please print the following text in the box below:

We, the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defense, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our posterity, do ordain and establish this CONSTITUTION for the United States of America.

We, the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defense, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our posterity, do ordain and establish this CONSTITUTION for the United States of America.

"Mr. Powell finds it easier to take it out of mothers, children and sick people than to take on this war industry," Mr. Brown commented daily. "Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

Mr. Powell finds it easier to take it out of mothers, children and sick people than to take on this war industry," Mr. Brown commented daily. "Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

Name: Alexander Debus

Pregled metoda

- Trening i validacioni skup
 - Stil pisanja svakog pojedinačnog karaktera
- Za svakog autora skupa za primenu:
 - Skup za prilagođavanje – istorija pisanja
 - Skup za prilagođavanje – alternativni klasifikatori i njihovi **vektori poverenja**
 - Skup za testiranje:
 - Primarno i alternativna predviđanja
 - Izbor **najpouzdanijeg**
 - Ažuriranje vektora poverenja

Klasterovanje stilova pisanja pojedinačnih karaktera

- Način poboljšanja (stil pisanja)?
- Motiv:
 - Konačan skup varijacija za 'a' (svi)
 - Dva razna 'a' (jedan)
- Skupovi slika sa istim labelama
- Trening i validacioni skup za bazni klasifikator

Klasterovanje stilova pisanja pojedinačnih karaktera

- Dobra reprezentacija karaktera?
 - Naivni pristup: slika
 - Naprediniji pristup? (novi prostor atributa, profinjen)
- Algoritam K sredina
 - $k = \min(30; 1 + \max(n/1000; 4))$
(evaluacija kvaliteta klasterovanja)
 - Težišta klastera – stil pisanja
 - Euklidska metrika


Kreiranje istorije pisanja

- Kada bazni CNN za trenutnog korisnika greši, i kako te greške ispraviti?
- Faza upotrebe modela (novi korisnici)
- Stil pisanja konkretnog karaktera od strane konkretnog korisnika
 - Imam stil pisanja karaktera
- Skup za prilagođavanje vs skup za testiranje

Kreiranje istorije pisanja

- Ulazna slika:
 - (ispravna labela, predviđanje CNNa)
 - Izlazi neurona preposlednjeg sloja baznog CNN
 - Najbliži klaster koji odgovara ispravnoj labeli
 - Interpretacija
 - Prosek* težišta klastera za svaki uređeni par

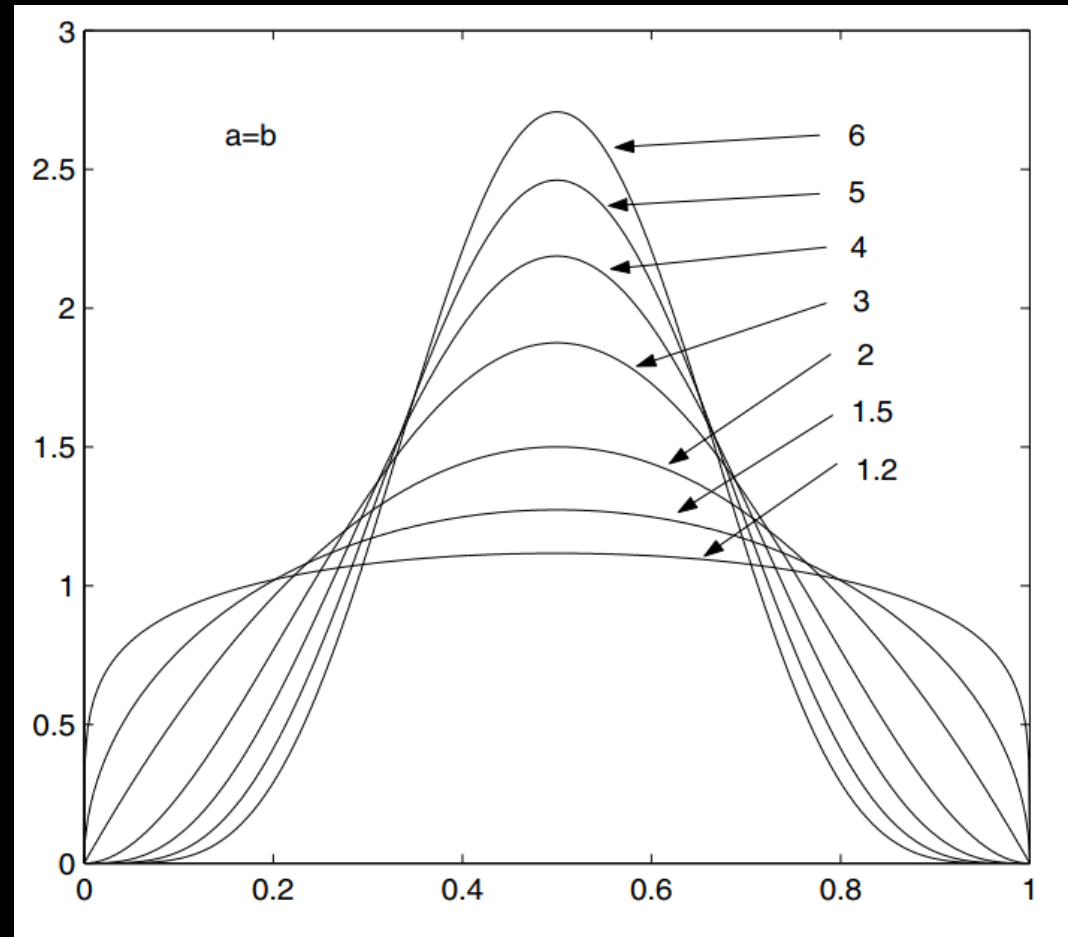
Upotreba istorije pisanja

- Upotreba upravo kreirane istorije pisanja
- Kada verovati baznom CNN a kada modifikovati njegova predviđanja i kako?
- Metod poboljšanja:
 - Fokus na baznu CNN
 - Alternativni klasifikatori metoda K najbližih suseda, za razno k
- Vektor poverenja baznog i alternativnih klasifikatora
 - 

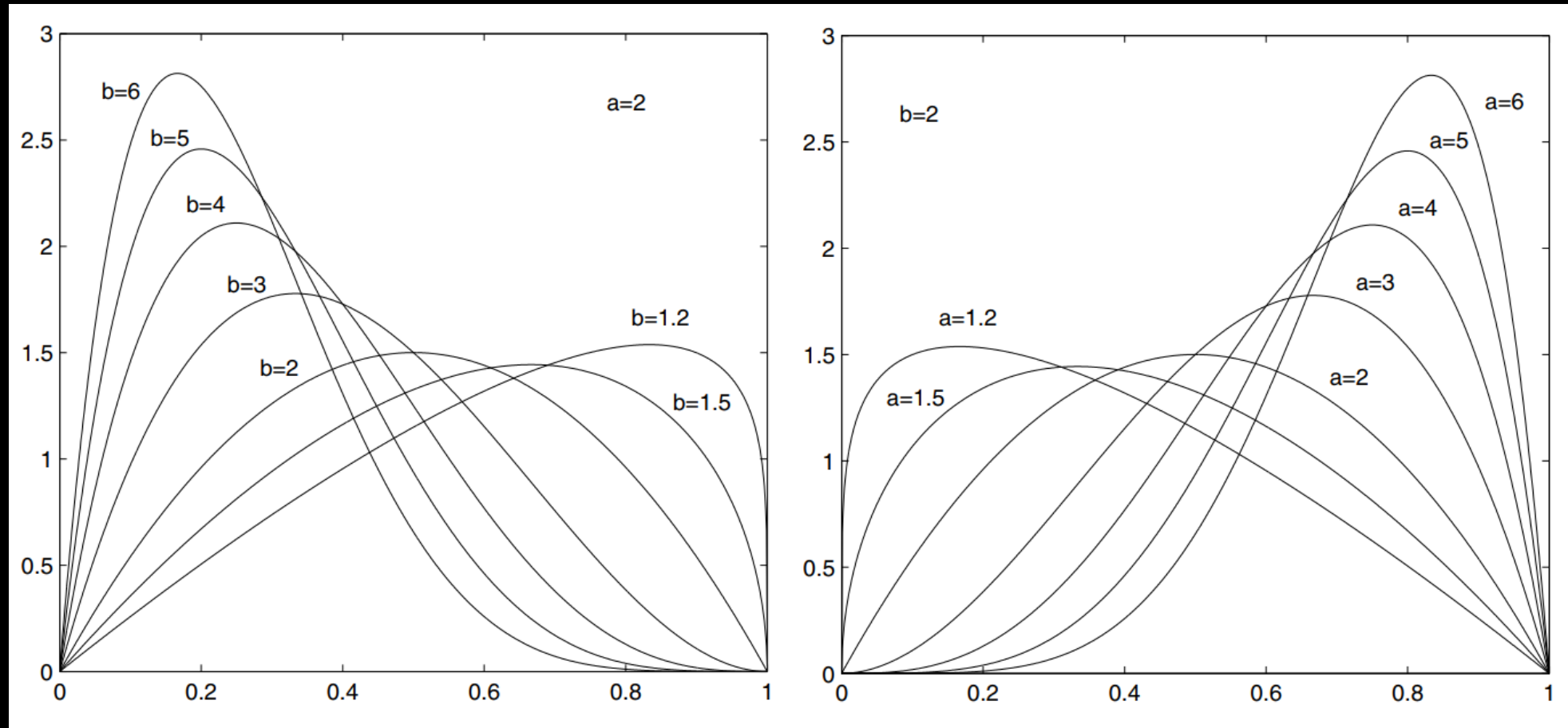
Vektor poverenja klasifikatora

- Niz očekivanja beta raspodela (62/70)
- Procena pouzdanosti klasifikatora prilikom predviđanja svake od labela
- Za svaku labelu kreira se po jedna beta raspodela
- Matematičko očekivanje
- Skup za prilagođavanje
 - Originalne labele, predviđanja
 - Realan scenario, ispravke

Vektor poverenja



Vektor poverenja



Upotreba istorije pisanja

- Simultana predviđanja
- Verujemo onom klasifikatoru koji za datu sliku ima najveću pouzdanost za labelu koju predviđa
- Konkretna slika – CNN predvidi l
- Korisnikova istorija pisanja
 - $(l, .*)$
 - (predviđena labela, ispravna labela)
 - Za svako k , nad tim vektorima pokrećemo KNN (u odnosu na ispravne labele)
- Vektor poverenja za svaki
- Ažuriranje parametara*

Evaluacija rešenja

- NIST Special Database 19: 2.3%-2.5%
 - 3600 autora, 62 labele, 225 slika/korisniku, 3.6 slika/korisnik x labela
- Deepwriting Dataset ETH Ciri: 2.7%
 - 294 autora, 70 labela, 1349 slika/korisniku, 19.2 slika/korisnik x labela
- Izuzetno mali resursi (knn)
- Bez retreniranja mreže (uređaji bez grafičkih karti)
- “Superskalabilnost”

Poređenje sa najboljim poznatim rezultatima

- Prevazilazi sve dosadašnje rezultate objavljene na NISTu (28x28)
- Neki od njih:
 - 2011. godine, ansambl CNN: 88.12%
 - 2012. godine, mreža sa više stubaca procesiranja: 88.37%
 - 2017. godine, CNN+SVM: 88.32%
 - 2018. godine, KNN+Random Forest: 75%
- 87.35% -> 89.60%

Poređenje sa najboljim poznatim rezultatima

- Ni jedan od prethodnih radova ne bavi se poboljšanjem klasifikatora oflajn rukom pisanog teksta
- Ni jedan od radova ne koristi direktno stil pisanja korisnika
- Na skupu podataka NIST ne postoje takvi radovi
- Na ETH Cirihi skupu podataka ne postoje objavljeni rezultati klasifikacije
- Postoje radovi koji se bave poboljšanjem klasifikatora oflajn rukom pisanog teksta
 - Nisu testirani na NIST/ETH
 - Ne uzimaju stil pisanja u obzir

Budući rad

- Transfer učenja (meta learning)
- Učenje na malim skupovima podataka (few-shot learning)
- Prilagođavanje trenutnom korisniku
- Small sample analiza:

Teorema o robusnosti (neprekidnosti); Elementarna varijanta

Teorema. Neka je M klasifikacioni model predstavljen neuronskom mrežom koji na konkretnom skupu podataka¹, označimo ga sa D , ostvaruje preciznost p , $p \in [0, 1]$. Neka je \bar{M} poboljšanje tog modela, kreirano na način opisan u radu, koje ostvaruje povećanje preciznosti $\Delta p > 0$. Tada za svaki klasifikacioni model N , predstavljen neuronskom mrežom, na istom skupu podataka ostvaruje preciznost p a njegovo poboljšanje \bar{N} kreirano na odgovarajući, opisan način, ostvaruje povećanje preciznosti Δp ako je model N " ϵ blizu" modelu M .

Drugim rečima, oko svakog modela čije poboljšanje ostvaruje neko, strogo veće od nule, povećanje preciznosti postoji " ϵ okolina" tako da za sve modele te okoline njihovo odgovarajuće poboljšanje takođe ostvaruje isto povećanje preciznosti.

milan_cugurovic@matf.bg.ac.rs