Random Forests

Matko Bošnjak, LIS, IRB

Sadržaj

- Uvod
- Random forest algoritam
- Svojstva
- Mogućnosti
- Nenadzirano učenje

Uvod

- Neke klasifikacijske/regresijske metode su nestabilne
  - male promjene u training setu ili proceduri konstrukcije modela uzrokuju velike promjene u predikciji (stabla odluke)
- Njihova se točnost može uvelike povećati perturbacijama i kombiniranjem
  - generiranje višestrukih inačica prediktora kroz perturbaciju training seta ili same procedure konstrukcije i kombiniranje u ansambl prediktora

Uvod

- Dobar ansambl prediktora
  - potrebno je izgraditi nekorelirane bazne prediktore
  - stabla odluke
    - niski bias, visoka varijanca (tj. nisu željeno precizna)
- Ansambl stabala odluke → "šuma stabala"
  - bias ostaje nizak, varijanca se smanjuje povećanjem broja baznih prediktora…(a korelacija?)
- Random forest
  - ansambl stabala odluke koji kombinira dva izvora "slučajnosti":
    - bootstrapping
    - metoda slučajnog potprostora

Izgradnja šume

- Bootstrapping
  - randomizacija podataka za treniranje
  - statistička metoda uzorkovanja s ponavljanjem
    - uzorkovanje N slučajnih instanci s ponavljanjem, rezultat dva skupa podataka
      - bootstrap skup – koristi se za izgradnju stabla
      - out-of-bag skup – koristi se za procjenu pogreške
  - radi se za svako stablo u šumi, time se osigurava da svako stablo uči na svojim podacima (niska korelacija)

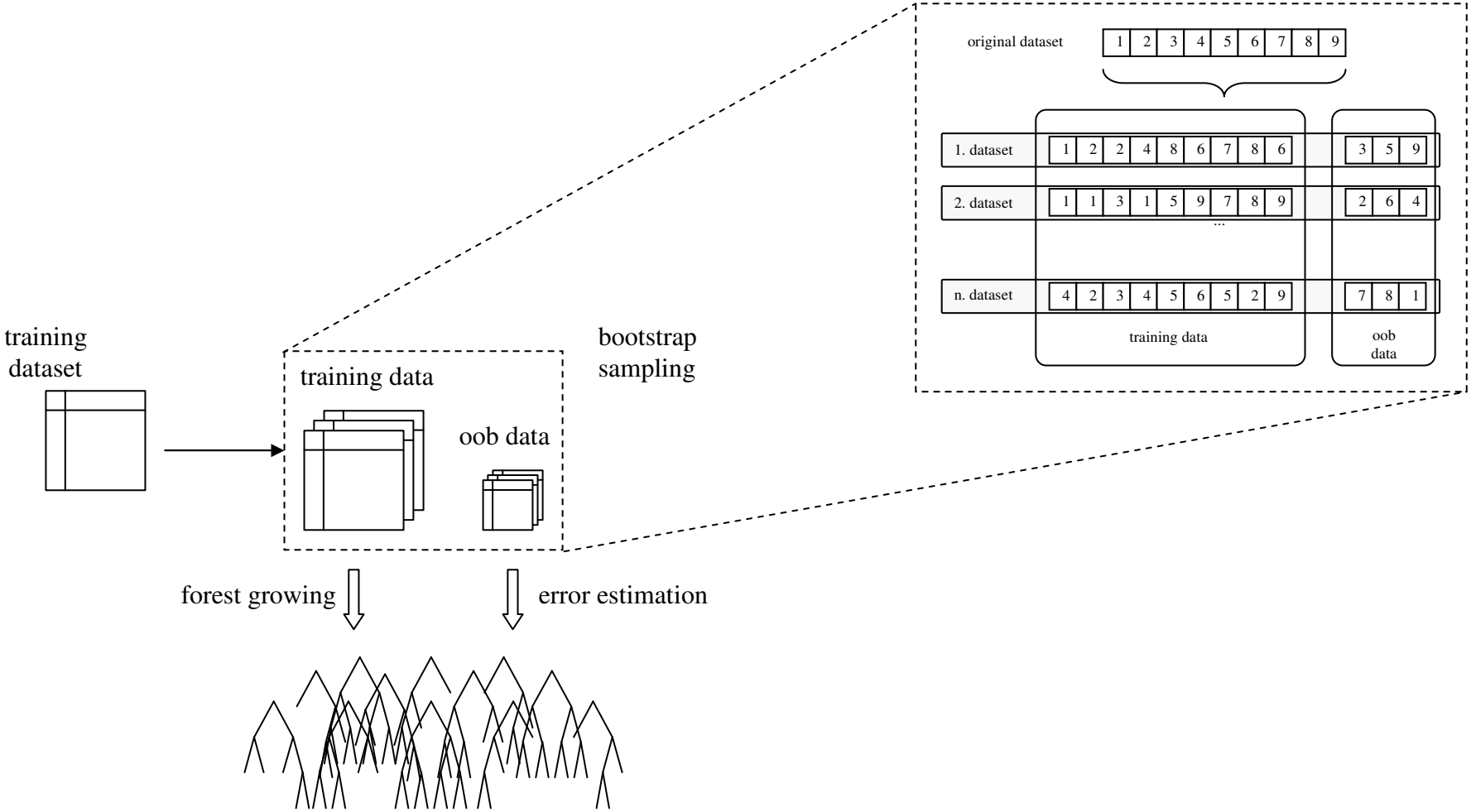
Izgradnja šume

- Izgradnja stabala
  - metoda slučajnog potprostora
    - slučajna selekcija atributa za vrijeme izgradnje stabala
    - odabir atributa na kojima se traži najbolji split
    - odabere se m slučajnih varijabli od mogućih M, nezavisno za svaki čvor
  - pronađe se najbolji split među tih m varijabli koristeći Gini index heuristiku za mjeru nečistoće
  - atribut s najvećim Gini indeksom je pobjednik – na njemu se radi split
  - stablo se gradi do maksimalne dubine, bez pruninga, da bi se maksimizirao bias

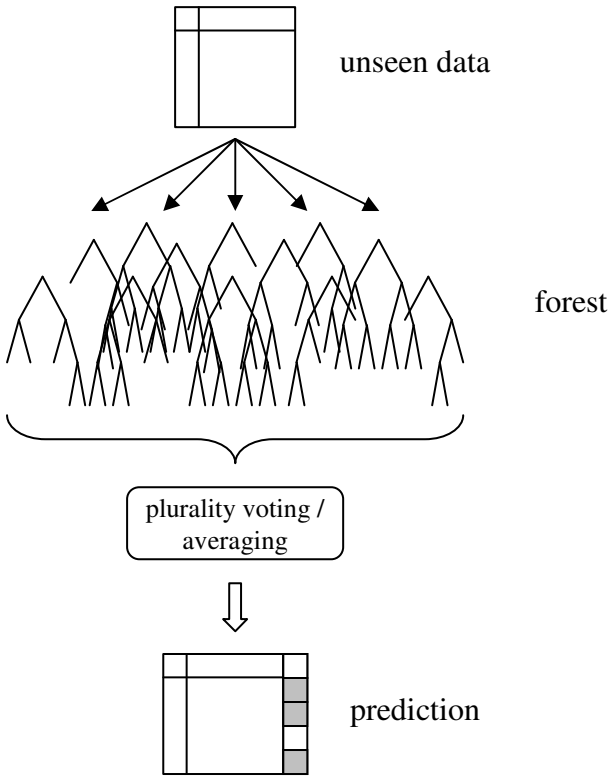
Izgradnja šume

- out-of-bag procjena pogreške
  - pri uzorkovanju, nisu svi podaci uključeni u konačni skup za učenje
  - oni koji nisu, zovu se out-of-bag (oob) podaci i koriste se u procjeni pogreške generalizacije generiranog stabla pa na kraju i čitave šume
  - oob procjena greške je proporcija pogrešne predikcije za pojedinu instance u oob skupu uprosječena po svim instancama
  - oob procjena dokazana kao nepristrana mjera
  - manji broj stabala → veća oob procjena pogreške
- Eliminira se potreba za zasebnim test setom

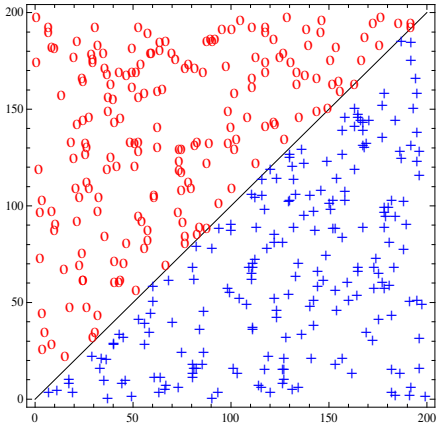
Izgradnja šume



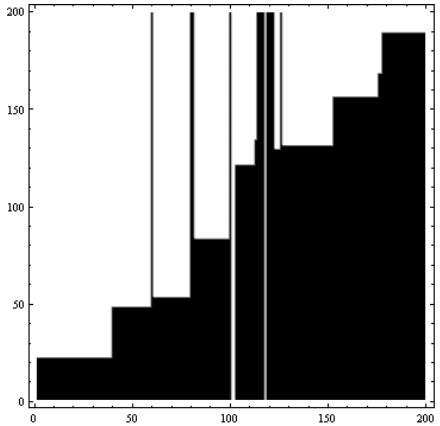
Korištenje šume



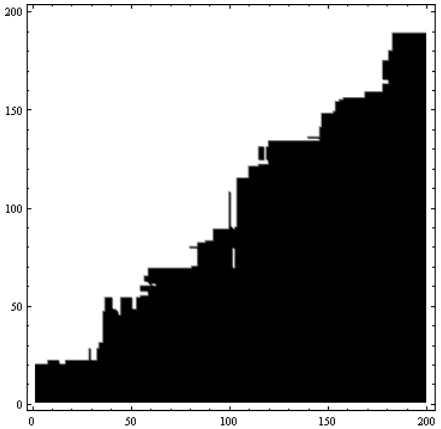
Kako to izgleda...



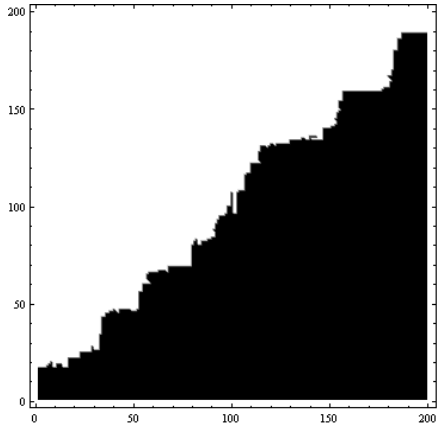
Umjetni podaci (200 instanci po podaci)
Linearna separacija, $y = x$



Model jednog stabla odluke
Greška 8,08%



Model 10 stabala
Greška 3,45%



Model 100 stabala
Greška 2,97%

Svojstva

- Svojstva
  - visoka preciznost
  - paralelizabilnost
  - sprječavanje overfittinga
  - kratko vrijeme izvođenja
  - omogućava obradu širokog spektra datasetova
  - nadomještanje nedostajućih vrijednosti

Svojstva

- Odlične prediktorske performanse
  - klasifikacija postiže bolje rezultate od regresije
  - postiže preciznost barem usporedivu s do sada najboljim poznatim ML metodama (AdaBoost i SVM)
  - visoka točnost prediktora (šume) se postiže brojem stabala
- Paralelna implementacija
  - trivijalna paralelizacija zadataka zbog trivijalne segmentacije u podzadatke (neovisnost podzadataka)
  - PARF (Paralelni RF Algoritam)
    - lako pokretanje na više procesora i/ili računala
    - Fortran 90 (Goran Topić, Tomislav Šmuc)
      - http://www.parf.irb.hr/
    - C# (Rajko Horvat)

Svojstva

- Overfitting
  - random forest eliminira problem overfittinga
  - zakon velikih brojeva
    - osigurava konvergenciju pogreške kada broj stabala $N \rightarrow \infty$
- Vrijeme izvođenja
  - brži od bagginga ili boostinga s istim brojem stabala (RF sa 100 stabala brži od AdaBoosta s 50)
  - RF sa 100 stabala (100 varijabli) jednako brz kao izgradnja 3 CART stabla
  - npr. sustav s 12 SPECint2006 bodova, izgradnja šume od 500 stabala, dataset od 11061 instanci i 42+1 atribut 3:15 min, šuma od 5000 stabala 36:50 min

Svojstva

- Efikasan u upravljanju širokim spektrom podataka
  - visokodimenzionalni podaci
  - podaci nepoznatih distribucija
  - podaci s nedostajućim vrijednostima
  - disbalansirani datasetovi
    - npr. podaci s KDDCup-a 2008, 127+1 atribut, izniman disbalans, 102294 instanci
    - RF omogućava umjetno balansiranje klasa kroz korisnički zadan skup težina po klasi
    - primjeri koji odgovaraju klasi na koju smo postavili težinu su adekvatno “utežani” u procesu izgradnje stabala

Svojstva

- RF nudi 3 različita rješenja za tretman nedostajućih vrijednosti
  - Zanemarivanje
    - nedostajuće vrijednosti se ne nadomještaju
  - Grubo nadomještanje
    - nadomjesci se izračunavaju pri prvom prolazu i ispunjavaju se prema tipu
      - kontinuirane – medijan svih ne-nedostajućih vrijednosti varijable u svojoj klasi
      - kategorijske – najfrekventnija vrijednost varijable u svojoj klasi
  - Namještanje temeljeno na “blizinama”
    - oslanja se na izračun “blizine” dvaju individualnih instanci
    - započinje s grubim nadomještanjem, nakon toga se izgradi šuma pa se izračuna matrica blizina parova instanci
    - nadomještanje utežano “blizinom” instanci

Mogućnosti

- Alati za pomoć pri interpretaciji modela
  - Matrica “blizina”
  - Značaj varijabli
  - Višedimenzionalno skaliranje
  - Prototipovi
  - Detekcija graničnih slučajeva (outliers)
  - Detekcija interakcija varijabli

Mogućnosti

- Matrica “blizina”
  - mjera sličnosti između parova instanci
  - jednaka je proporciji stabala za koja dvije različite instance završe u istom listu (terminalnom čvoru) stabla
- Značaj varijabli
  - koristi se za uvid u strukturu podataka i sam proces predikcije
  - definira se kao doprinos preciznosti predikcije – prediktivna moć varijable
  - RF implementira dva različita načina izračuna značaja varijabli
    - brzi značaj varijabli
      - baziran na činjenici da split na čvoru specifične varijable smanjuje Gini indeks
    - permutacijski značaj varijabli
      - bazirana na razmjeru pogrešne klasifikacije originalnih i podataka sa permutiranim vrijednostima promatrane varijable

Mogućnosti

- Višedimenzionalno skaliranje
  - matrica blizina individualnih instanci se može iskoristiti za dobivanje kvadrata udaljenosti u euklidskom prostoru
  - te udaljenosti se onda mogu koristiti za metričko skaliranje, aproksimaciju vektorskog prostora u niže dimenzije pogodne za grafičku reprezentaciju
- Prototipovi
  - umjetno konstruirane instance koji se mogu koristiti za zaključivanje o relaciji varijable i same predikcije
  - oni su vrsta reprezentativnih primjera odgovarajuće klase

Mogućnosti

- Detekcija outliera
  - nepodobne instance, udaljene od distribucije svoje klase
  - malena blizina između instance i ostalih instanci indicira visoku mjeru nepodobnosti
- Detekcija interakcija varijabli
  - dvije (nezavisne) varijable su u interakciji ukoliko split na jednoj varijabli čini split na drugoj sistematski manje ili više mogućim
  - temeljene na Gini indeksu

Nenadzirani RF

- Nenadzirano učenje
  - problem strojnog učenja u kojemu primjeri nemaju eksplicitnu oznaku klase
  - cilj je pronaći uzorke, odrediti strukturu
  - problem: nema mjera koju je potrebno optimizirati
- RF se može koristiti za nenadzirano učenje
  - originalni podaci – klasa 1
  - sintetski podaci – klasa 2
    - permutirani originalni podaci
  - umjetno stvoren problem dvaju klasa
    - može se procesirati RF-om koristeći sve navedene interpretacijske alate
    - MDS od posebnog značaja – daje uvid u podatke
  - oob procjena pogreške modela <50% → međuzavisnosti među varijablama

Zaključak

- Svestrana, precizna metoda koja pruža mnoštvo alata za interpretaciju

…i samo čeka na primjene ☺

Hvala na pažnji