Katedra za informatiku, Fakultet tehničkih nauka Novi Sad

Soft computing

dr Đorđe Obradović

dr Vuk Malbaša

#14

Učenje sa obeležjima

- Često je mnogo jeftinije praviti predikcije nego ručno obeležavati podatke
- Veliki skupovi podataka bez obelezja se lako mogu napraviti, značajno ih je teže obeležiti
- Ideja: Hajde da pitamo algoritam koji podaci su mu najzanimljiviji?
  - Na kojim podacima je algoritam naj sigurniji? Na kojima ne?
  - Moguća domenska razlika, ono što nas interesuje nije isto što je zanimljivo algoritmu

Aktivno učenje

- Aktivno učenje je skup pristupa gde algoritam koji uči neko preslikavanje f(x)=y, a ujedno bira primere x iz kojih uči.
  - Pošto se pojedinačno biraju primeri često je mnogo manje podataka potrebno za uspešno treniranje
- Najvažnije pitanje, koji podaci su bitni za treniranje?
  - Primeri na kojima algoritam greši
  - Primeri na kojima je predikcija algoritma nesigurna

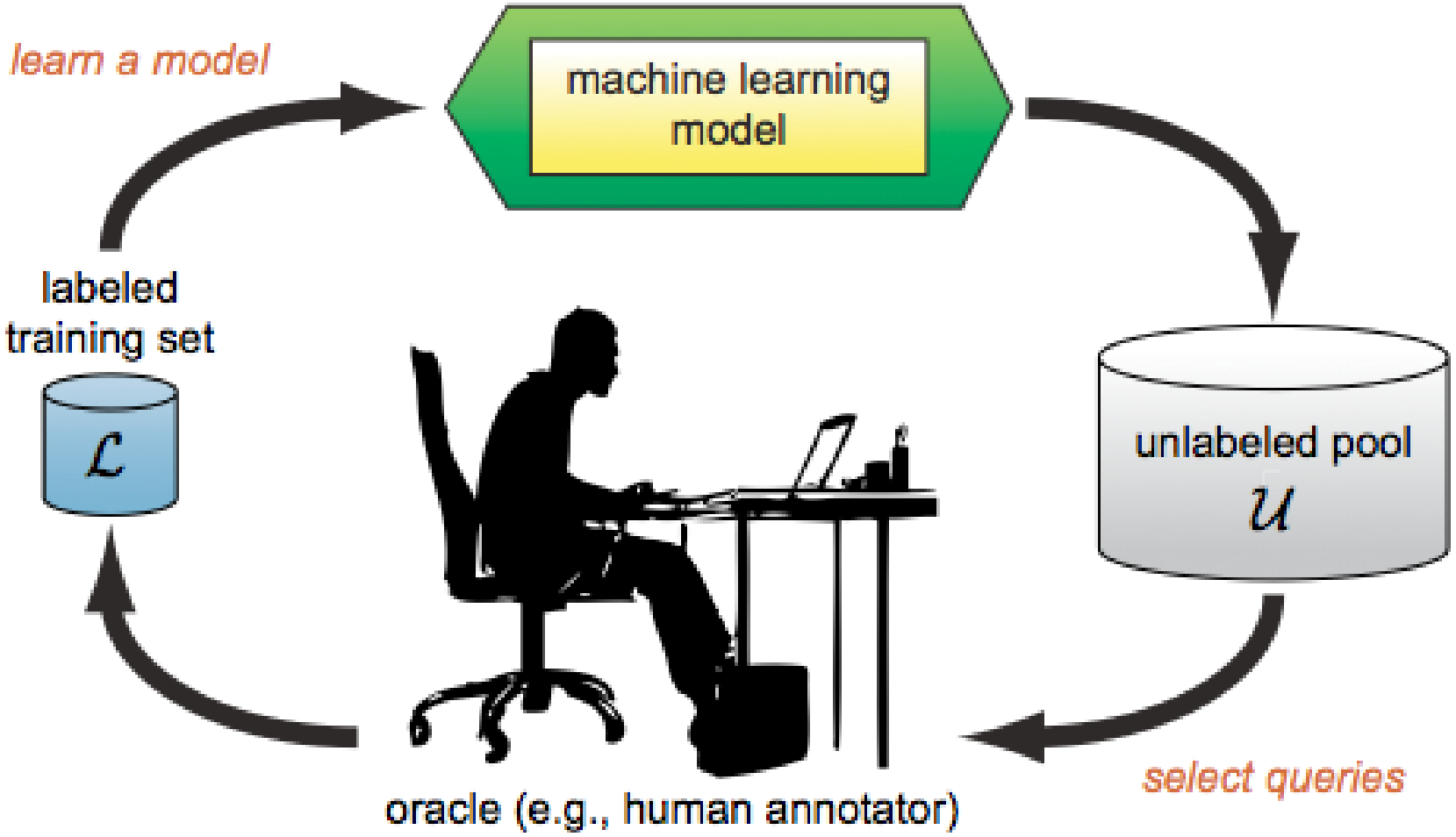
Biranje podataka za treniranje

- Nekoliko različitih primena u zavisnosti od modela koji koristimo
- Često se koriste pristupi bazirani na sigurnosti predikcije
  - Ako su labele iz skupa {-1,1} tada je sigurnost u predvidjanje udaljenost od granice odluke f(x)= 0
    - U ovom slučaju je dovoljno gledati samo magnitudu predikcije.
    - Kod linearnog modela su moguće sve vrednosti dok su predviđanja pa samim tim i sigurnost ograničena za logistički model
  - Može se interpretirati kroz verovatnoću podataka x za koji je p(y|x) najbliže 0.5
    - Tada je verovatnoća obeležja je najmanje sigurna.

Pool based algoritam

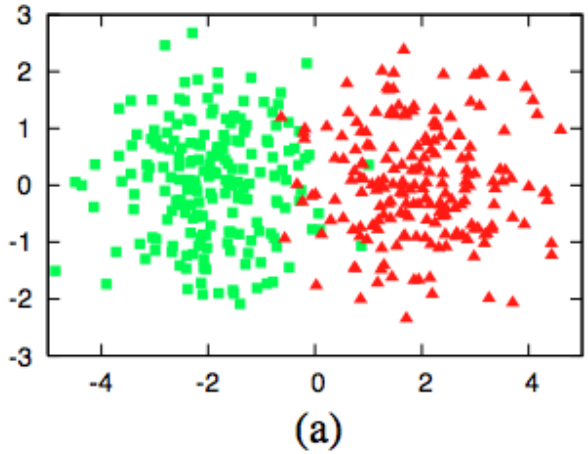
- Inicijalizuj U, L
  - U su ne obeleženi podaci, u početku su svi podaci u skupu U
  - L su obeleženi podaci, u početku je skup L sadrži
    - Inicijalizacija L na po jedan primer iz svake klase
    - Random inicijalizacija L
- while(nismo ispunili budžet, postigli tačnost, ...)
  - Istreniraj se na L
  - Napravi predikcije na U
  - Izaberi primer x* iz U na kojem je predikcija namanje sigurna
  - Ručno označi x* i ukljuci ga u L a izbaci iz U

Pool based active learning

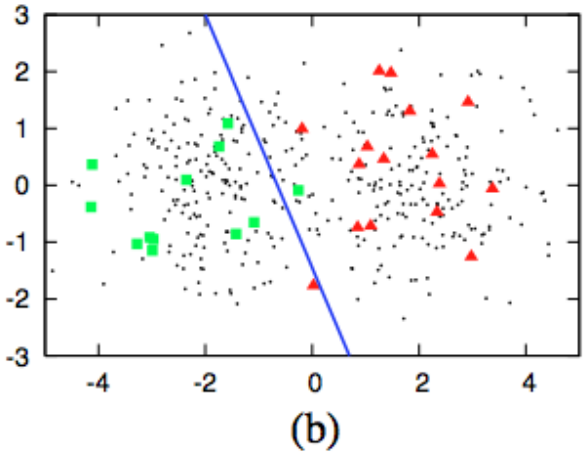


[Settles 2010]

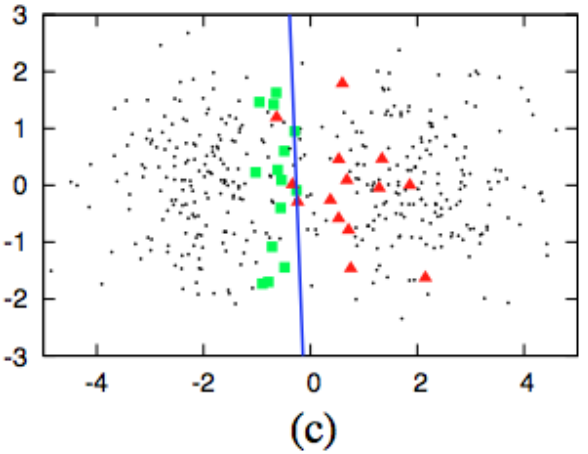
Primer: Linearna regresija



Obeležja svih trening podataka



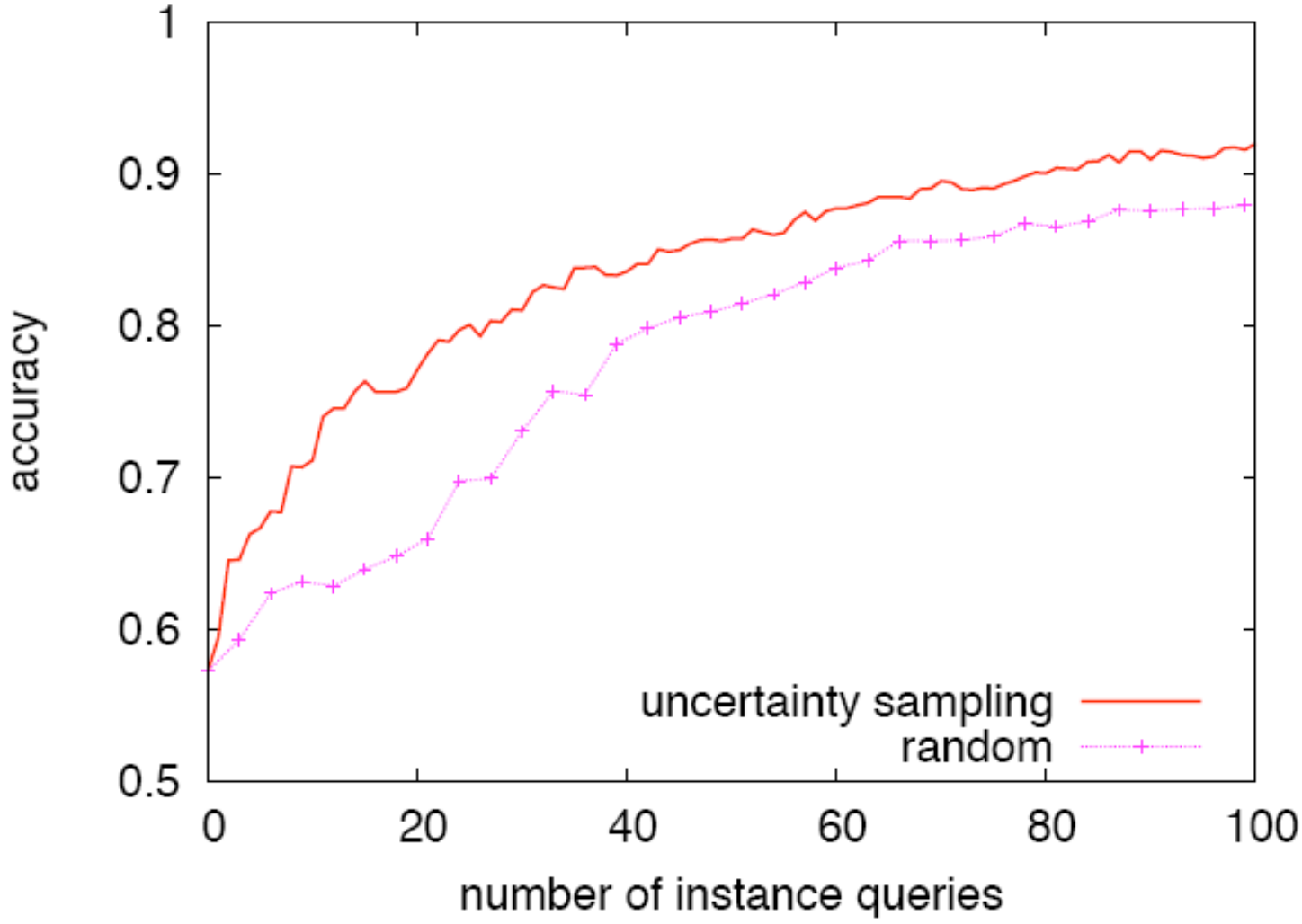
Random izbor 70% tačnost posle 30 obeleženih primera



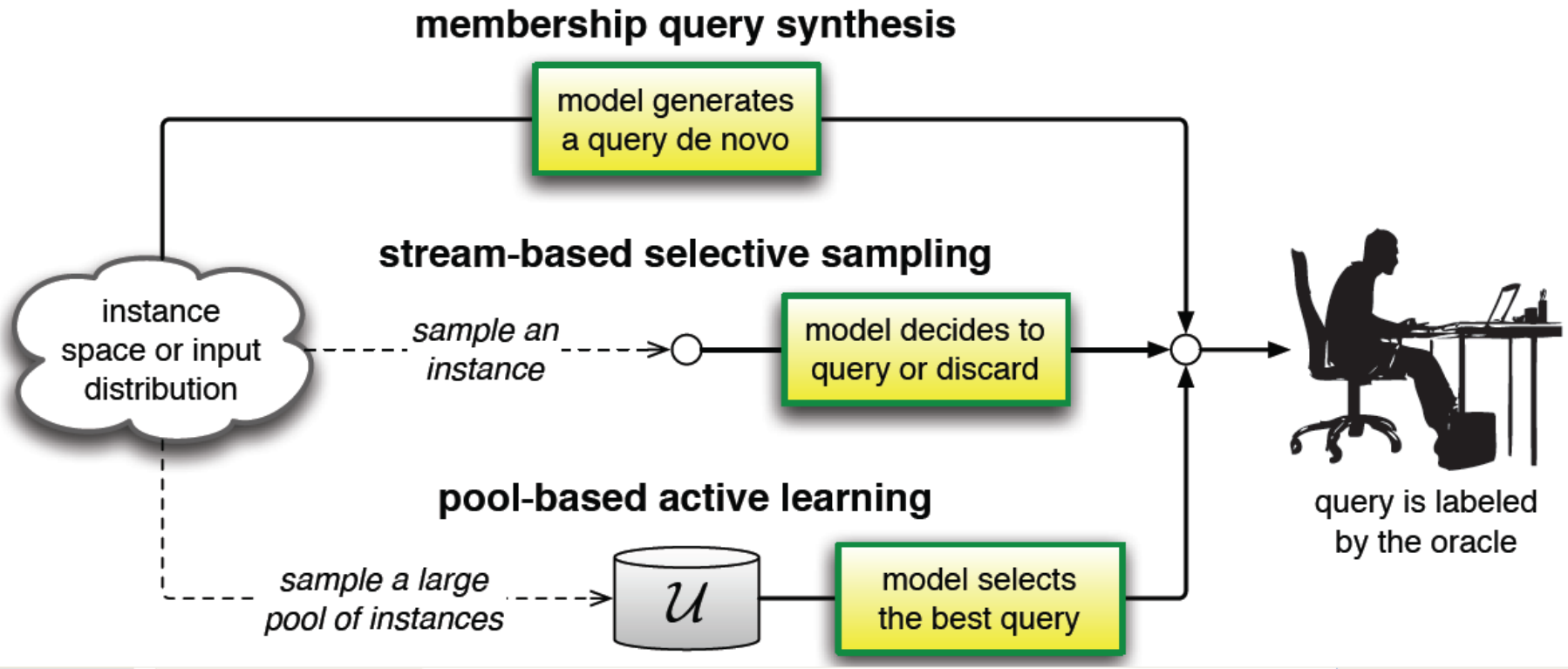
Active learning

- Sa malim brojem interesantnih primera uspeli smo značanjno da popravimo tačnost predviđanja

Performanse



Alternativni pristupi



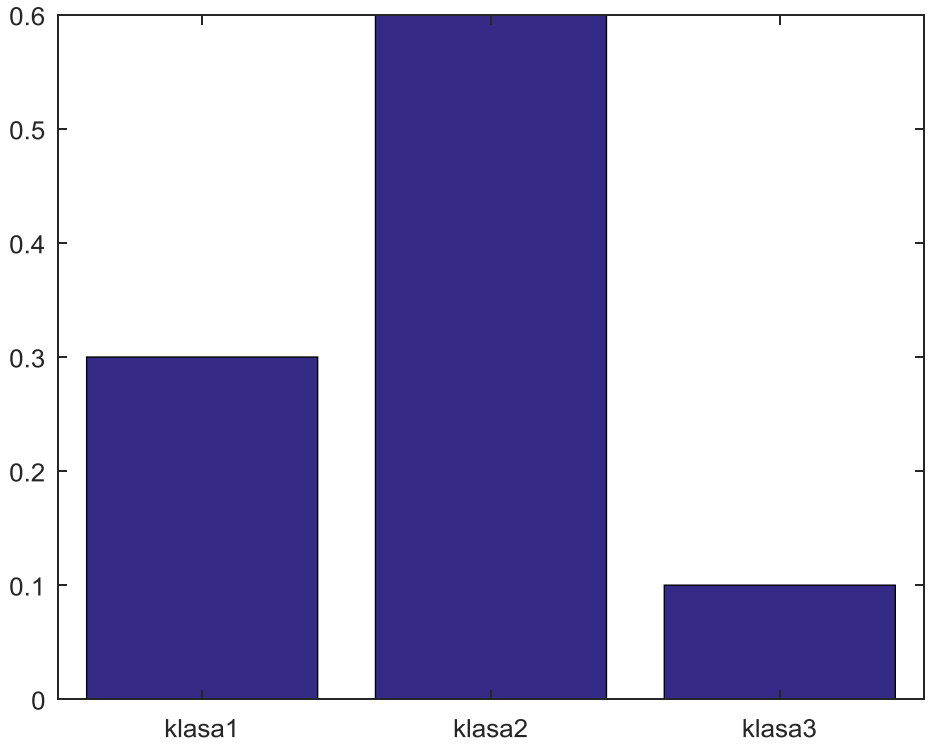
- Burr Settles 2009

Alternativni pristupi

- Membership query synthesis
  - Novi primer se generišu a ne biraju od postojećeg skupa
  - Mora postojati način da se generišu novi primeri što ponekad nije trivijalno
- Stream based selective sampling
  - Potrebno je odrediti koji podaci se čuvaju pošto ih ima previše da bi ih čuvali sve

Multi class problem

- Šta kada imamo više od jedne klase?
- Možemo problem svesti na binarni ako koristimo tri modela
- Tada dobijamo verovatnoću predikcije za sve klase



Multi class problem

- Oduzmimo od verovatnoće najverovatnije klase verovatnoću druge najverovatnije klase
- Kada je prediktor siguran tada ce biti velika razlika između najveovatnije i druge po redu
- Kada prediktor nije siguran tada će sve klase imati od prilike istu verovatnoću

Problemi

- Sparse podaci
- Nedostajuće vrednosti
- Nebalansirane klase
- Kategoričke primenljive
- Šum u podacima