Projekti na predmetu: Strojno učenje – PMF (2011)

Tomislav Šmuc; Matko Bošnjak

Uvod
Kao projekte u okviru ovog predmeta predlažemo učešće na nekom od tzv."challenge"-a u području data mining-a odnosno machine learning-a.
Primarno je to DM Cup – koji je i namijenjen studentima. KDDCup su tipično zahtjevni problemi, pa je to ovdje više kao informacija
Mjesta na kojima možete naći druge challenge i realne probleme su TunedIT i Kaggle
Prijedlog je da se organizirate u ekipe od po 3 člana. Možete izabrati bilo koji od challenge-a (bilo u tijeku ili završenih) ili dolje predloženih challenge-a, ali pazite na vremenska ograničenja. Dobro proučite uvjete natjecanja. Predviđamo da će vam za dolaženje do (relativno kvalitetnih) rješenja trebati minimalno 2-3 tjedna.

Datumi:
a) formiranje ekipa i odabir zadatka – najbolje do 15.04. (e-mail), najkasnije prijava na predavanju 18.04.
b) predaja izvještaja 23.05.2011.
c) prezentacije: 23.05.2011., (25.05.) 30.05.2011.

Za one koji ne žele sudjelovati u „challenge" projektu – ostavljamo mogućnost da naprave predavanje o posebnim područjima primjene algoritama strojnog učenja (vidi pod III i IV).

I) Međunarodna natjecanja vezana uz data mining

DATA-MINING-CUP

Izazov:

Početak: 15.04.2010.

Rok za slanje: 31.05.2010.

Nagradni fond:
Ako osvojite jedno od prva tri mjesta, također imate osiguranu ocjenu iz predmeta. Dodatni bodovi će se dodjeljivati u ovisnosti o vašoj poziciji na ukupnoj rang listi.

Uvjeti: Sudionici moraju biti isključivo studenti. **Samo dvije grupe s jednog fakulteta (provjeriti)** !

Detalji na: http://www.data-mining-cup.de/en/dmc-competition/

Izvještaj: Uz finalne rezultate morate poslati i kratki izvještaj o tome kako ste rješavali problem. Taj izvještaj je nužan da biste bili prikazani na konačnoj rang listi (i uvjet je za ocjenu vašeg projekta).
Na engleskom jeziku, u obliku članka.

KDD- Cup 2011

Izazov: **Recommending Music Items based on the Yahoo! Music Dataset**

Kraj challenge-a: 30.06.2011

Detalji na: http://www.kdd.org/kdd2011/kddcup.shtml

VL.Net challenge

Izazov: **Recommending lectures**

Rok za slanje: 31.05.2010. (kraj challenge-a: June 30)

Detalji na: TunedIT

NAPOMENE: Ovo su natjecanja sa različitim profilima: od početnika (većina) do iskusnih znalaca u svom području. Neka vas loš plasman ne zabrinjava, a odličan plasman neka vam svakako laska (dodatni bodovi!).

Potrebno je pridržavati se pravila propisanih odgovarajućim izazovom, sastavi timova su proizvoljni. Dozvoljeno je sve, svaka nova ideja za rješavanje zadatka. Kreativnost se dodatno nagrađuje.

Dodatne informacije i konzultacije:

- e-mail (tomislav.smuc@irb.hr; matko.bošnjak@irb.hr)
- u sklopu termina vježbi

II) Realizacija algoritama/metoda, primjene na posebne probleme (modifikacija postojećeg algoritma u R-u, Weka-i, Rapid Miner-u ...) (ekipa: 2-3 člana)

- Robustni k-nn (+ određivanje težine atributa, detekcija outlier-a)
  Modifikacija postojećeg k-nn algoritma ili vlastita realizacija.

  Literatura: PAGER: Parameterless, Accurate, Generic, Efficient kNN-based Regression, A. Desai et al., IIIT/TR/2009/157.

- Co-training algoritam za polu-nadzirano učenje R

- Q-learning (implementacija uz dinamičko generiranje primjera); Vlastiti algoritam

- Multi-label classification=> Instalacija i povezivanje MULAN+WEKA (MEKA)

- Preporučivanje/ocjenjivanje filmova – realizacija kolaborativnog filtering algoritma za NetFlix tip podataka ? (R – recommenderlab)

Projekti vezani uz VL.Net dataset

- Primjeniti metode nenadziranog (unsupervised) učenja na grafovima za clustering grafa definiranog „lecture co-viewing" frekvencijama. Koristiti ostale podatke u analizi: autore, taksonomiju, event-e.

III) Predavanja o posebnim temama (1-2 člana)

Oni koji odaberu ovu opciju, trebaju napraviti predavanje (mi možemo ponuditi pomoć u odabiru literature) o jednoj temi, prema izboru. Predviđeno je da ta predavanja budu u sklopu termina predviđenog za predavanja – polovicom i krajem mjeseca svibnja. Ovo su naši prijedlozi:

a) **Strojno učenje u obradi teksta:** software, problemi u obradi prirodnog teksta, i algoritmi za obradu prirodnog jezika (NLP + text mining), GATE, U-compare
b) **Strojno učenje u obradi slika i video zapisa**: software, problemi u obradi slika, i algoritmi (Open Vision)
c) **Učenje u stalnom dotoku podataka** (datastreams algoritmi) (Patrik Đurđević ?)
d) Sustavi za predlaganje/preporučivanje (Recommender systems & collaborative filtering)
e) **Učenje sa složenom strukturom ciljne varijable** (multi-label, hijerarhija kategorija, ontologija)
f) **Kako učiti rangirati** (info retrieval)?
g) Kako radi Watson; Kako uči NELL ?
h) **Duboke arhitekture:** Deep encoders + Restricted Boltzmann machines

IV) Vlastiti odabir i definiranje projekta ili teme u dogovoru s prof/asistentom (ekipa: zavisi o složenosti problema)

- Text mining: otkrivanje plagijata u korpusu tekstova (?)
- Klasifikacijski problemi u analizi slika (?)