Projekti na predmetu: Strojno učenje – PMF (2012)

Tomislav Šmuc; Ivan Ivek, Tomislav Lipić

Uvod

Cilj projekata je da praktično realizirate neke od tehnika strojnog učenja koje ćete upoznati na predmetu Strojno učenje (može se desiti i da koristite i neke nove tehnike koje nisu pokrivene predmetom) na nekom od realnih problema, te da svoj projekt znate opisati i prezentirati kao malo znanstveno istraživanje u području.

Datumi:

a) formiranje ekipa i odabir zadatka – najbolje do 19.03. (e-mail), najkasnije prijava na predavanju 21.03.
b) „kontrolna prezentacija“ – 25.4. ili 2.5.
c) predaja izvještaja 16.05.2012.
d) prezentacije (tentativno): 23.05.2012. i 30.05.2012.

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

I. Aktualna natjecanja vezana uz data mining (ekipa 2-3 člana)

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. Otvoreni challenge-i su tipično zahtjevni problemi. Mjesta na kojima možete naći druge challenge i realne probleme su TunedIT, Kaggle i Pascal2. Prijedlog je da se organizirate u ekipe od po 3 člana. Možete izabrati bilo koji od dolje predloženih challenge-a (ili neki drugi koji je u tijeku ili završen), ali pazite na vremenska ograničenja. Dobro proučite uvjete natjecanja. Predviđamo da će vam za dolaženje do (relativno kvalitetnih) rješenja trebati minimalno 2-3 tjedna.

1. DATA-MINING-CUP (DMC)

- Registracija za DMC natjecanje počinje: **1.3.2012**
- DMC natjecanje počinje: **3.4.2012**
- Rok za predaju rezultata: **15.5.2012**
- 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 ovaj uvjet)** !
- 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.
  - Izvještaj je nužan da biste bili prikazani na konačnoj rang listi (i uvjet je za ocjenu vašeg projekta).
  - Na engleskom jeziku, u obliku članka.

2. Benchmark Bond Trade Price Challenge (@Kaggle)

- Završava: 30.04.2012. (problem se može rješavati i nakon završetka natjecanja)
- Detalji na: http://www.kaggle.com/
- **NAPOMENE:**
  - Ovo je otvoreno natjecanje i mogu se natjecati svi: 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; ivan.ivek@irb.hr)
  - u sklopu termina vježbi

II. Mali istraživački projekti

Projekti istraživačkog karaktera predstavljaju nove ideje/metode za rješavanje nekog problema koje je potrebno evaluirati/validirati/istražiti. Ideje projekta su razrađene, te je zadatak studenta njihovo ostvarenje. Rezultat projekta se potencijalno može objaviti u nekom znanstvenom časopisu ili konferenciji.

1. **(Hierachical) multilabel classification with output sparsity assumptions as a pooling design (Group testing)**
   - Opis:
     - Pronalazak kodova za kompresiju više labela (stabla) u sažeti prostor na temelju pretpostavke rijetkosti originalnog prostora
     - Instalacija i povezivanje MULAN+WEKA (MEKA)
   - **Dodatne informacije i konzultacije**:
     - e-mail (ivan.ivek@irb.hr)
     - u sklopu termina vježbi

2. **Correlation-based Subset Selection in Multilayered Group Method of Data Handling**
   - Opis:
     - Samoorganizirajuće polinomske mreže
     - Weka implementacija - wGMDH: http://wgmdh.irb.hr/en/project
     - Evaluacija nove varijante algoritma za učenej GMDH modela
     - Implementacija optimizacije modela (gradijentni spust, Levenberg-Marquardt)
   - **Dodatne informacije i konzultacije**:
     - e-mail (ivan.ivek@irb.hr)
     - u sklopu termina vježbi

III. Realizacija algoritama/metoda i/ili primjene na posebne probleme (ekipa: 2-3 člana)

Ova skupina projekata uključuje istraživanje i tehničku implementaciju metoda za rješavanje trenutno aktualnih tema u području.

1. **Clustering za izrazito velike+rijetke matrice (online clustering; Recommender problemi)**
2. **Time-series prediction pipeline (Yahoo-finance historical dana, Zg burza) (download-transform + Rapid Miner/R workflow)**
3. **Music retrieval pipeline**
4. **Self-training & Co-training algoritam za polu-nadzirano učenje realiziran u Rapid Miner-u ili R-u**

IV. Ostale moguće teme i problemi

U okviru ove skupine predlažemo neke dobro poznate probleme dubinske analize podataka (*data mining*) i primjene metoda strojnog učenja, kao i aktualne istraživačke teme iz područja. Opširan popis skupova podataka prikladnih za neke dolje spomenute teme/probleme može se naći na:

- http://kevinchai.net/datasets
- http://www.datawrangling.com/some-datasets-available-on-the-web
- http://theinfo.org/

1. Problemi iz *Challenges in Machine Learning*: http://www.chalearn.org/

Od studenta se očekuje da izabere, prouči i pristupi rješavanju jednog od problema predstavljenih u CHALERN izazovima prethodnih godina. Uz svaki izazov može se pronaći službena stranica na kojoj se nalazi detaljni opis problema i pripadni skup podataka (*dataset*). Za svaki izazov objavljeno je i specijalno izdanje u časopisu *Journal of Machine Learning Research* u kojem se mogu naći objavljeni članci o pristupu rješavanja pojedenih izazova. Student ne treba raditi iscrpnu pretragu i analizu postojećih rezultata te je dovoljno proučiti 2-3 članka koji ga mogu usmjeriti i pomoći pri rješavanju problema odabranog izazova.

Popis izazova nalazi se ovdje: http://www.chalearn.org/challenges.html

Primjerice student može izabrati:

- *Feature Selection Challenge* iz 2003, http://www.nipsfsc.ecs.soton.ac.uk/datasets/
- *Gasture Challenge* aktualni izazov iz ove godine, http://gesture.chalearn.org/
    - Paziti nema još objavljenih rezultata koji mogu pomoći kako smjernica za rješavanje
- Ili bilo koji drugi koji mu se čini zanimljiv sa popisa: http://www.chalearn.org/challenges.html

2. Problemi vezani za otkrivanje znanja u skupovima podataka (*Data mining; DM*)

Cilj projekta je analizirati jedan stvarni skup podataka vezan za određeni problem. Ova tema uključuje i usporedbu različitih metoda strojno učenja na jednom odabranom specifičnom problemu. Jedan od mogućih izvora skupova podataka (problema):

- Data Mining Datasets:
  http://www.inf.ed.ac.uk/teaching/courses/dme/2012/datasets.html
  Ovdje se za svaki skup podataka nalazi pripadni opis problema. Većinom su to problemi klasifikacije ili klasteriranja. Za svaki problem priloženi su opisi prijašnjih odabranih metoda koje mogu pomoći pri rješavanju odabranog problema.
- KDnuggets: http://www.kdnuggets.com/datasets/index.html
    - UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
- Analiza podataka iz nekog *DM challange*a (nije potrebno prisustvovati natjecanju)
    - Npr. Mining Challenge - Android platforme, http://2012.msrconf.org/challenge.php

3. **KDD Cup 2012 - User Modeling based on Microblog Data and Search Click Data, http://www.kddcup2012.org/**

Cilj projekta je proučiti probleme (Track 1 i Track 2) KDD Cup 2012, izabrati jedan problem (može i oba) i pristupiti njegovu rješavanju. Studentu se preporuča da dodatno prouči tematiku problema ili javi asistentima za usmjerenje.

4. **Modeliranje i analiza društveni mreža i društvenih medija (vijesti, blogovi, twitter)**

Student umjesto rješavanja jednog od problema na KDD Cup 2012 natjecanju može izabrati proizvoljan problem koji je vezan za analizu društvenih medija i društvenih mreža. Studentu se preporuča da prouči pogleda Jure Leskovca kako bi se zainteresirao za problem:

- Jure Leskovec, http://cs.stanford.edu/people/jure/
- KDD 2011 Tutorial - Social Media Analytics, http://snap.stanford.edu/proj/socmedia-kdd/
- Effects of User Similarity in Social Media, http://cs.stanford.edu/people/jure/pubs/similarity-wsdm12.pdf
- Modeling Social and Information Networks: Opportunities for Machine Learning, http://videolectures.net/icml09_leskovec_msain/
- Skupovi podataka:
    - http://www.trustlet.org/wiki/Datasets
    - http://www.trustlet.org/wiki/Repositories_of_datasets
    - http://www.trustlet.org/wiki/Trust_network_datasets#Released_datasets

5. **Otkrivanje mišljenja i analiza sentimenta iz tekstualnih podataka (Opinion Mining and Sentiment Analysis)**

Studenta zainteresiranog za ovu temu upućujemo na materijale:

- Opinion Mining and Sentiment Analysis, http://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf
    - 1. Poglavlje daje dobar uvod u problem
    - 7. Poglavlje izlistava izvore skupova podataka koji se mogu koristiti pri izboru ove teme
- Sentiment Analysis and Subjectivity, http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf
- Diplomski rad, http://www.doc.ic.ac.uk/teaching/distinguished-projects/2011/l.carstens.pdf
- Estimating Sentiment Orientation in Social Media for Intelligence Monitoring and Analysis, http://ngc.sandia.gov/assets/documents/SentimentAnalysisSocialMedia_IEEEISI2010.pdf
    - Podaci: http://www.cs.cornell.edu/People/pabo/movie-review-data/
- Neka druga tema vezana uz obradu prirodnog jezika (natural language processing)
    - Jedan od alata - MALLET, http://mallet.cs.umass.edu/
    - Naći/Istražiti/Proučiti ostale alate za NLP
    - Skupovi podataka:
        - http://research.microsoft.com/nlp/
        - http://nlp.stanford.edu/links/statnlp.html
        - http://trec.nist.gov/data/reuters/reuters.html

6. Sustavi preporučivanja (Recommender Systems)

Cilj ove vrste projekata je dobro proučiti i istražiti temu vezanu za sustave preporučivanja. U konačnici student bi trebao rekonstruirati i implementirati neki opisani algoritam vezan za sustave preporučivanja. Studenta se upućuje na sljedeće materijale:

- Stanford ml-class - XVI. RECOMMENDER SYSTEMS, http://www.ml-class.org/course/video/preview_list - Lagani uvod u sustave preporučivanja
- Poglavlje iz knjige *Mining of Massive Datasets*: http://i.stanford.edu/~ullman/mmds/ch9.pdf
- Recommender Systems Tutorial (KDD'10): http://pages.cs.wisc.edu/~beechung/kdd10-tutorial/index.html
- Deepak Agarwal, Recommender Systems: The Art and Science of Matching Items to Users, http://www.research.rutgers.edu/~eliassi/mlseminar.html#Deepak
- Deepak Agarwal, http://research.yahoo.com/Deepak_K_Agarwal
    - fLDA: Matrix Factorization through Latent Dirichlet Allocation, http://research.yahoo.com/pub/3051
- Yehuda Koren, Collaborative Filtering with Temporal Dynamics, http://research.yahoo.com/files/kdd-fp074-koren.pdf
    - Opisuje algoritam za preporučivanje koji uključuje informaciju o vremenskoj ovisnosti
- Improving Recommendation Accuracy by Clustering Social Networks with Trust, http://www.cs.umd.edu/~tdubois/publications/trustClusters.pdf
    - FilmTrust dataset, http://trust.mindswap.org/FilmTrust/about.shtml
- Preporučivanja članaka vijesti (*News articles recommander systems*)
    - CHU, WEI, http://www.gatsby.ucl.ac.uk/~chuwei/
        - Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms
        - A contextual-bandit approach to personalized news article recommendation
- Preporučivanje ljudi (*People to people recommander systems*)
    - CCR-A Content-Collaborative Reciprocal Recommender for Online Dating, http://users.cis.fiu.edu/~lli003/recsys/papers/reciprocal/3672.pdf
    - A people-to-people recommendation system using Tensor Space Models, http://eprints.qut.edu.au/47886/1/ACM_SAC_Sangeetha_eprints.pdf
    - Finding someone you will like and who won't reject you, http://sydney.edu.au/engineering/it/~irena/umap2011.pdf
    - Learning to Make Social Recommendations: A Model-Based Approach, http://aminer.org/PDF/adma2011/7121/71210124.pdf
    - Stochastic matching and collaborative filtering to recommend people to people, http://users.cis.fiu.edu/~lli003/recsys/papers/reciprocal/PizzatoSilvestriniRecSys2011.pdf
    - Learning Collaborative Filtering and Its Application to People to People Recommendation in Social Networks, http://www.cse.unsw.edu.au/~wobcke/papers/learning-to-rank.pdf
    - A Hybrid Content-Collaborative Reciprocal Recommender for Online Dating, http://sydney.edu.au/engineering/it/research/tr/tr667.pdf
    - People Recommendation Based on Aggregated Bidirectional Intentions in Social Network Site, http://www.cse.unsw.edu.au/~wobcke/papers/rules.pdf

- Collaborative Filtering for People to People Recommendation in Social Networks, http://www.cse.unsw.edu.au/~wobcke/papers/bilateral-recommendation.pdf

7. Zaključivanje o povjerenju u društevnim mrežama

- Network Clustering Approximation Algorithm Using One Pass Black Box Sampling, http://www.cs.umd.edu/~tdubois/fast-clustering.pdf
  - Klasteriranje zasnovano na povezanim komponentama u podgrafu slučajno induciranih bridova
- Predicting Trust and Distrust in Social Networks, http://www.cs.umd.edu/~golbeck/papers/sign.pdf
- Skupovi podataka:
  - http://www.trustlet.org/wiki/Datasets
  - http://www.trustlet.org/wiki/Repositories_of_datasets
  - http://www.trustlet.org/wiki/Trust_network_datasets#Released_datasets
  - http://www.trustlet.org/wiki/Repositories_of_datasets

8. Detekcija i prepoznavanje ljudi i lica na slikama (Human and face detection and recognition)

Primjer projekta ove tome može uključivati sustav koji će za danu sliku lica osobe provjeriti koliko je pouzdana dana slika. Sustav možete zamisliti kroz:

1. Provjeru je li slika sadrži jedno lice čovjeka (*human and face detection*)
2. Provjeru je li detektirano lice na slici odgovara nekim drugim licima iz baze korisnika (face recognition)

- Izvor informacija: http://www.facedetection.com/
- Viola Jones face detection framework – prvi sustav za detekciju lica u realnom vremenu , http://research.microsoft.com/en-us/um/people/viola/Pubs/Detect/violaJones_CVPR2001.pdf
- Face Recognition by Humans: 20 Results all Computer Vision Researchers Should Know About, http://web.mit.edu/bcs/sinha/papers/20Results_2005.pdf
- Algoritmi za prepoznavanje lica, http://www.face-rec.org/algorithms/#Image
- Face Recognition via Sparse Representation, http://perception.csl.uiuc.edu/recognition/Home.html
- Detekcija ljudi na slici, http://www.merl.com/reports/docs/TR2006-068.pdf
- Popis skupova podataka:
  - http://vision.ai.uiuc.edu/mhyang/face-detection-survey.html#face-database
  - http://www.facedetection.com/facedetection/datasets.htm
  - http://www.face-rec.org/databases/
  - http://www.idiap.ch/resource/frontalfaces/
  - http://vis-www.cs.umass.edu/lfw/

9. Podaci i ideje vezane uz Microsoft Kinect: http://en.wikipedia.org/wiki/Kinect

- Detekcija aktivnosti ljudi (Human Activity Detection) pomoću Microsoft Kinecta:
  - http://pr.cs.cornell.edu/humanactivities/index.php
  - http://pr.cs.cornell.edu/papers/human_activity_detection_rgbd_2011.pdf
  - http://pr.cs.cornell.edu/papers/unstructured_human_activity_learning.pdf
- Gasture Challenge, http://gesture.chalearn.org/
- NYU Depth Dataset, http://cs.nyu.edu/~silberman/site/?page_id=27

    - Indoor Scene Segmentation using a Structured Light Sensor, http://cs.nyu.edu/~silberman/papers/indoor_seg_struct_light.pdf
- UW's RGB-D Object Dataset, http://www.cs.washington.edu/rgbd-dataset/
    - A Scalable Tree-based Approach for Joint Object and Pose Recognition, http://www.cs.washington.edu/homes/kevinlai/publications/lai_aaai11.pdf
- B3DO: Berkeley 3-D Object Dataset, http://kinectdata.com/
- Semantic Structure From Motion (SSFM): http://www.eecs.umich.edu/vision/projects/ssfm/index.html

10. Prepoznavanje aktivnosti (Activity recognition)

Studenta zainteresiranog za ovu temu upućujemo na materijale:

- Tutorial on Human Activity Recognition, http://cvrc.ece.utexas.edu/mryoo/cvpr2011tutorial/
- Skup podataka za prepoznavanje aktivnosti ljudi, https://sites.google.com/site/tim0306/datasets
    - Preporuča se pogledati *Benchmark datasets* i pripadni članak
- Activity Recognition: Datasets, Bibliography and others, http://www.cse.ust.hk/~derekhh/ActivityRecognition/index.html
- Opportunity project
    - Skup podataka, http://www.opportunity-project.eu/challengedatasetdownload
    - Publikacije, http://www.opportunity-project.eu/node/56

11. Dubinska analiza nizova podatka i vremenske serije (Data Streams Data Mining, timeseries analysis)

- Poglavlje iz knjige *Mining of Massive Datasets*: http://i.stanford.edu/~ullman/mmds/ch4.pdf
- Mining Data Streams: A Review, http://www.sigmod.org/publications/sigmod-record/0506/p18-survey-gaber.pdf
- Data Stream Mining – A Practical Approach, http://www.cs.waikato.ac.nz/~abifet/MOA/StreamMining.pdf
- Skupovi podataka i materijali za klasifikaciju i klasteriranje vremenskih serija
    - Eamonn Keogh datasets, http://www.cs.ucr.edu/~eamonn/time_series_data/
    - http://www.cs.ucr.edu/~eamonn/tutorials.html
    - Time Series Data Library, http://robjhyndman.com/TSDL/

12. Neuronske mreže i Boltzman machines

- Hinton
    - http://www.cs.toronto.edu/~hinton/csc321/index.html
    - http://www.cs.toronto.edu/~hinton/csc2535/index.html
    - http://www.utstat.toronto.edu/~rsalakhu/code.html
- Neural Networks framework
    - http://www.heatonresearch.com/encog

V. Proizvoljni vlastiti projektni prijedlog

Student može predložiti vlastiti proizvoljni projektni prijedlog. Projektni prijedlog može se zasnovati i na temama iz IV). Kako bi što lakše i potpunije opisali svoju ideju za projekt pokušajete u njegovoj razradi odgovoriti na sljedeća pitanja, napravljena po uzoru na Heilmeierov kriterij:

- Koji je problem koji pokušavate riješiti?
  - Definirajte ciljeve svoga problema.
  - Koji skup podataka koristite.
- Koje su postojeće metode kojim se problem rješavao i koji su nedostaci postojećih metoda.
- Na koji način će te probati riješiti problem?
  - Koje metode/algoritme/tehnike mislite koristiti?
  - Opišite to što specifičnije možete
- Kako mislite ocijeniti uspješnost rezultata svoga projekta?
- Što očekujete predati kako konačni rezultat projekta?