Strojno učenje – PMF (2018)
Dodatne teme za projekte

U ovom dokumentu predloženo je nekoliko znanstveno-istraživačkih tema za studentske projekte koje su vezane za područja istraživanja na Institutu Ruđer Bošković. Namjenjene su prije svega visokomotiviranim studentskim timovima i studentima koji su dugoročno zainteresirani za istraživački rad. Navedene teme mogu poslužiti i kao polazna točka za definiranje teme diplomskog rada. Organizacija projektnog zadatka, način njegova izvođenja i rokovi opisani u dokumentu „Upute za provođenje projektnih zadataka“ (na stranici kolegija Projektni zadaci) jednaki su za sve studentske timove bez obzira odabere li tim jednu od navedenih tema ili predloži svoju vlastitu temu.

Tema 1. Primjena strojnog učenja u području biologije prema sljedećoj motivaciji:

U mnogim klasifikacijskim problemima iz područja biologije na ulazu se koristi informacija o sastavu proteinske sekvence. Sekvencijski podaci predstavljaju izazov za standardne algoritme strojnog učenja koji na ulazu očekuju vektor fiksne duljine. U slučaju proteinske sekvence tipično se konstruiraju značajke koje predstavljaju sumarnu statistiku sekvence (npr. frekvencije aminokiselina) ili njena fizičko/kemijska svojstva. Međutim, problem s ovako konstuiranim značajkama je da se gubi informacija o slijedu aminokiselina u sekvenci. U Asgari et al. [1] se problem učenja reprezentacije sekvence proteina rješava koristeći tehnike iz obrade prirodnog jezika pri čemu 3-grami reprezentiraju riječ, a sekvenca proteina rečenicu. Nakon dobivanja reprezentacije 3-grama koristeći word2vec metodu, dobivene reprezentacije se sumiraju i predstavljaju reprezentaciju sekvence.

Zadatak 1.1. Sistematična provjera optimalne duljine n-grama proteinskih sekvenci

Iako autori u radu [1] navode da su ispitali n-grame duljine 3-6 koristeći k-NN, sistematična usporedba nije napravljena. U ovom radu cilj je na sistematičnoj usporedbi n-grama duljine 3-6 (ili 8 ovisno o broju ostvarivih kombinacija). Potrebno je trenirati ~500k proteinskih sekvenci koristeći word2vec i dobiti reprezentacije n- grama. Zatim se dobivene reprezentacije n-grama sumiraju i koriste za rješavanje klasifikacijskog problema u biologiji (npr. predikcija klase proteina, termofilnosti).

Dodatni zadatak: Osim word2vec reprezentacija, zbog sistematične usporedbe radi se i usporedba koristeći PCA i Glove reprezentacije.

Zadatak 1.2. Sistematična provjera optimalnog načina agregacije embeddinga proteinskih sekvenci

U ovom zadatku cilj je na sistematičnoj provjeri dobivanja vektorske reprezentacije sekvence. Potrebno je uzeti naučene reprezentacije 3-grama dobivenih word2vec modelom iz [1] te isprobati razne načine agregacije (slično kao u radu [2]) uključujući: 1) suma; 2) min/max; 3) suma nad normaliziranim vektorima; 4) min/max nad normaliziranim vektorima; 5) 1&2 ili 3&4 (ovisno koji ostvaruje bolje rezultate); 6) tf-idf weighting sume; 7) SIF reweighting [3]. Kvaliteta metode agregacije provjerava se na nezavisnim problemima kao što su predviđanje familije proteina, te predviđanje termofilnosti.

Dodatni zadatak: Osim word2vec reprezentacija, zbog sistematične usporedbe radi se i usporedba koristeći PCA i Glove reprezentacije.

Zadatak 1.3. Korištenje konvolucijskih neuronskih nreža za dobivanje reprezentacije proteinske sekvence

Problem kod sumiranja reprezentacija 3-grama u radu [1], ponovno se gubi informacija o slijedu i globalnom kontekstu. U ovom radu je cilj je dobiti reprezentaciju sekvence koristeći konvolucijske neuronske mreže. Mreža će se trenirati na umjetno konstruiranom problemu pri čemu je cilj dobiti reprezentaciju. Dobivena reprezentacija se zatim provjerava na nezavisnim problemima kao što su predviđanje familije proteina, te predviđanje termofilnosti. Prijašnji radovi pokazali su uspješnost konvolucijskih neuronskih mreža za predviđanje proteinske strukture [2,3,4].

---

[1] Asgari, Ehsaneddin, and Mohammad RK Mofrad. "Continuous distributed representation of biological sequences for deep proteomics and genomics."PloS One10.11 (2015): e0141287.

[2] De Boom, Cedric, et al. "Large-scale user modeling with recurrent neural networks for music discovery on multiple time scales." Multimedia Tools and Applications (2017): 1-23

[3] Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings." (2016).

[4] Hou, Jie, Badri Adhikari, and Jianlin Cheng. "DeepSF: deep convolutional neural network for mapping protein sequences to folds." *Bioinformatics* (2017).

[5] Wang, Sheng, et al. "Protein secondary structure prediction using deep convolutional neural fields."*Scientific reports* 6 (2016): 18962.

[6] Zhou, Jian, and Olga G. Troyanskaya. "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction." *arXiv preprint arXiv:1403.1347*(2014).

Tema 2. Napredna analiza vremenskih nizova velikih razmjera

Projektni zadataci u sklopu ove teme imaju visoki znanstveno istraživački karakter strukturiran kroz dvije osnovne cjeline. Prva cjelina uključuje razvoj te kvantitativno i kvalitativno (interpretacija) vrednovanje metodologije za analizu (s fokusom na grupiranje i klasifikaciju) velikog skupa vremenskih nizova[1] (TS). Metodologija objedinjava (1) individualne TS vektorske reprezentacije zasnovane na velikom skupu predodređenih specifičnih značajki[2,3,4,5]; (2) individualne TS vektorske reprezentacije ugrađene u niže dimenzionalni prostor pogodne za vizualizaciju[6,7]; (3) naučene individualne TS vektorske reprezentacije zasnovane na dubokom učenju[8,9] i (4) izvedene značajke iz međuodnosnih (graf) TS reprezentacija[10] zasnovanih na modeliranju (tj. zaključivanju o strukturi[11,12,13]) i analizi kompleksnih mreža[14,15]. Poseban naglasak biti će na evaluciji značajki inspiriranih područjem kompleksnih sustava uključujući mjere komplektnosti (npr. approximate entropy, sample entropy). Druga cjelina zadatka uključuje specifičnu primjenu razvijene metodologije s fokusom na interno dostupni podatkovni skup aktivacija neurona miševa prilikom izvođenja određene radnje. Kroz tu primjenu pokušat će se ustanoviti specifični obrasci memorije i učenja na mreži aktivacija nurona. Ostale primjene uključuju područje fiziologije čovjeka, financija i ekosustava kriptovaluta.

Sva potrebna literatura citirana je u danom sadržaju projektnog zadatka.

Alati:

- Omogućiti će se pristup uspostavljenoj programskoj okolini i računalnoj infrastrukturi (pristup preko Jupytera)
- TSFRESH: Time Series Feature extraction based on scalable hypothesis tests, https://github.com/blue-yonder/tsfresh
- HCTSA: highly comparative time-series analysis, https://github.com/benfulcher/hctsa
- TIGRAMITE – Causal discovery for time series datasets, https://jakobrunge.github.io/tigramite/
- m-TSNE: https://github.com/minhnguyen-cs/mtsne; LION-TSNE: https://github.com/andreyboytsov/LION-tSNE
- Stanford Network Analysis Platform (SNAP), https://github.com/snap-stanford/snap; graph-tool, https://graph-tool.skewed.de/performance

Podatkovni skupovi:

- Interni skupovi podataka (područje neuroznanosti, fiziologije)
- https://physionet.org/; https://www.synapse.org/#!Synapse:syn8717496/wiki/422884

---

[1] Esling, Philippe, and Carlos Agon. "Time-series data mining." *ACM Computing Surveys (CSUR)* 45.1 (2012): 12.

[2] Fulcher, B. D. (2017). Feature-based time-series analysis. arXiv preprint arXiv:1709.08055.

[3] Christ, Maximilian, Andreas W. Kempa-Liehr, and Michael Feindt. "Distributed and parallel time series feature extraction for industrial big data applications." arXiv preprint arXiv:1610.07717 (2016).

[4] Fulcher, Ben D., and Nick S. Jones. "Highly comparative feature-based time-series classification." IEEE Transactions on Knowledge and Data Engineering 26.12 (2014): 3026-3037. + Fulcher, Ben D., Max A. Little, and Nick S. Jones. Highly comparative time-series analysis: the empirical structure of time series and their methods. Journal of The Royal Society Interface 10.83 (2013): 20130048.

[5] Schäfer, Patrick, and Ulf Leser. "Fast and accurate time series classification with weasel." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017. + Schäfer, Patrick, and Ulf Leser. "Multivariate Time Series Classification with WEASEL+ MUSE." arXiv preprint arXiv:1711.11343 (2017).

[6] Boytsov, A., Fouquet, F., Hartmann, T., & LeTraon, Y. (2017). Visualizing and Exploring Dynamic High-Dimensional Datasets with LION-tSNE. arXiv preprint arXiv:1708.04983.

[7] Nguyen, M., Purushotham, S., To, H., & Shahabi, C. (2017). m-TSNE: A Framework for Visualizing High-Dimensional Multivariate Time Series. arXiv preprint arXiv:1708.07942.

[8] Malhotra, Pankaj, et al."TimeNet: Pre-trained deep recurrent neural network for time series classification." aarXiv:1706.08838 (2017).

[9] Gamboa, John Cristian Borges. "Deep Learning for Time-Series Analysis." arXiv preprint arXiv:1701.01887 (2017).

[10] Ferreira, Leonardo N.& Liang Zhao. Time series clustering via community detection in networks.Information Sciences 326 (2016): 227-242.

[11] Runge, J., Sejdinovic, D., & Flaxman, S. (2017). Detecting causal associations in large nonlinear time series datasets. arXiv:1702.07007.

[12] Wang, Wen-Xu, Ying-Cheng Lai, and Celso Grebogi. "Data based identification and prediction of nonlinear and complex dynamical systems." Physics Reports 644 (2016): 1-76.

[13] Hallac, David, et al. "Network Inference via the Time-Varying Graphical Lasso." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017.

[14] Benson, A. R., Gleich, D. F., & Leskovec, J. (2016). Higher-order organization of complex networks. Science, 353(6295), 163-166.

[15] https://github.com/thunlp/NRLPapers

Tema 3. Detekcija lažnih vijesti

Zadataj 3.1. News Stance Detection

Zadatak je za par (naslov, tekst) detektirati odnos teksta prema naslovu: slaganje, neslaganje, diskusija (tekst samo diskutira naslov), nepovezanost [16]. Ovakav klasifikator se može koristiti u većem pipeline-u za detekciju lažnih vijesti. Dostupan je dataset i kodovi tri najbolja modela na challenge-u. Zadatak se može podjeliti u tri podzadatka, bazični je nužan minimum, i uključuje reprodukciju rezultata za challenge-a za tri najbolja modela. Trebalo bi još napraviti i dio analize na kakvim primjerima modeli griješe, te gdje se međusobno slažu ili razlikuju. Dodatni zadatak je napraviti novi model, nadogradnjom najboljih sa challenge-a ili razvojem novog modela. Posebno interesantno bi bilo iskoristiti SoA textual entailment model [17] za feature extraction. Može se probati i korištenje vanjskih baza znanja (wikipedia) za obogaćivanje zadataka u naslovu, te razvoj nekog novog modela, deep modela ili neke dobre kombinacije feature extractora + klasifikatora.

Zadatak 3.2. Detekcija istinitosti izjava

LIAR dataset [18] sastoji se od kratkih izjava klasificiranih u različite stupnjeve istinitosti. U članku razmatrani klasifikatori predviđaju istinitost samo na temelju teksta i rade očekivano jako loše pošto je teško provjeriti istinitost činjenica bez znanja o svijetu. Bazični zadatak bi bio replicirati neki od najboljih klasifikatora iz članka, napraviti analizu grešaka. Dodatni zadatak bi bio koristiti vanjsku bazu (više manje) istinitih činjenica (Wikipedia, ...) kako bi se poboljšale performanse. Također je moguće iskoristiti SoA textual entailment model [19] npr. provjerom da li neka pouzdano točna izjava povlači izjavu čija istinitost se utvrđuje.

Tema 4. Neural Topic Models Evaluation

Topic Modeli su modeli za eksplorativnu analizu tema [20]. Unatrag nekoliko godina pojavili su se tzv. Neuralni Topic Modeli. Zadaci su sljedeći (odabrati podskup od ponuđenog). (1) implementacija modela [21]. (2) upogonjivanje ili proširivanje modela [22]. Model je originalno napravljen kao fuzija language modela i topic modela. Bilo bi jako dobro izdvojiti i učiniti samostojećim samo topic-model dio, ako ne samo pokrenuti postojeći [23]. (3) evaluacija modela. Koji kod pristup se odbere, nakon što se uspije istrenirati dobar model, treba odabrati neki veći skup tekstova (Wikipedia ili njen podskup, NYT, ...), istrenirati model i zatim pogledati teme. Metode evaluacije su sljedeće. (i) kvalitativna analiza. Pogledati 100 ili više tema (pogledati top riječi i top dokumente za teme), probati ih labelirati sa konceptima, vidjeti kakav uvid u strukturu zbirke tekstova daju, vidjeti koliko je tema dobro (odgovara konceptima) a koliko je šum. (ii) analiza koherentnosti tema. Uzeti gotov kod za računanje koherentnosti tema [24]. Mjeriti koherentnost neuralnog modela te standardnog LDA modela (npr. gensim ili mallet implementacije) koristeći nekoliko mjera koherentosti sa najboljim performansama. (iii) feature extraction. Uzeti neki klasifikacijski dataset (20Newsgroup, ...) te koristiti Neural-TM za feature extraction te mjeriti performanse klasifikacije u odnosu na LDA.

---

[16] http://www.fakenewschallenge.org/
[17] https://github.com/lukecq1231/enc_nli/
[18] https://arxiv.org/pdf/1705.00648.pdf
[19] https://github.com/lukecq1231/enc_nli/
[20] https://dl.acm.org/citation.cfm?id=2133826
[21] https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewFile/9303/9544
[22] https://arxiv.org/pdf/1704.08012.pdf
[23] https://github.com/jhlau/topically-driven-language-model
[24] https://github.com/dice-group/Palmetto, https://github.com/jhlau/topic_interpretability