PD.TVZ.HR
POLYTECHNIC & DESIGN Vol. 12, No. 1, 2024.
Mrežna inačica: ISSN 2459-6302
Tiskana inačica: ISSN 1849-1995
DOI: 10.19279/TVZ.PD.2024-12-1-06

KONCEPTUALNA DISKUSIJA O OBJAŠNJIVOSTI NADZIRANOG UČENJA ZNAČAJKI ZA KLASIFIKACIJU

CONCEPTUAL DISCUSSION OF EXPLAINABILITY OF SUPERVISED FEATURE LEARNING FOR CLASSIFICATION

Dino Vlahek[1], Bojan Nožica[2]

[1] UM FERI, Koroška cesta 46, 2000 Maribor
[2] Tehničko veleučilište u Zagrebu, Vrbik 8, Zagreb, Hrvatska

SAŽETAK

U ovom radu predstavljene su osnovne ideje nadziranog učenja značajki za klasifikaciju. Posebna pozornost pridaje se objašnjivosti tih pristupa. Metode učenja značajki su neobjašnjive ili ograničene u svojim predikacijskim rezultatima što je posljedica nemogućnosti rekombiniranja ulaznih značajki. Pristupi koji omogućuju povećanje dimenzionalnosti prostora ulaznih značajki su spori jer zahtijevaju iterativne nekonveksne optimizacije i podešavanje brojnih konfiguracija skrivenih dimenzija. U tim slučajevima autori uglavnom ne daju objašnjenja naučenog modela. Međutim, objašnjenja se mogu postići s različitim stupnjevima uspjeha s učenjem interpretativnih modela oko danog uzorka od interesa ili procjenom važnosti svake značajke u rezultatu klasifikacije.

Ključne riječi: *objašnjiva umjetna inteligencija, klasifikacija, učenje značajki, otkrivanje znanja*

ABSTRACT

This paper presents the basic ideas of supervised feature learning for classification. Special attention is given to the explainability of these approaches. Feature learning methods are either inexplicable or limited in their prediction results due to the inability to recombine input features. Approaches that increase the dimensionality of the input feature space are slow because they require iterative non-convex optimizations and tuning of numerous configurations of hidden dimensions. In these cases, authors generally do not provide explanation of the learned model. However, explanations can be achieved in various degrees of success by learning interpretive models around a given pattern of interest or by evaluating the importance of each feature in the classification output.

Keywords: *explainable artificial intelligence, classification, feature learning, knowledge discovery*

1. UVOD
1. INTRODUCTION

Značajka (engl. Feature) se definira kao individualna karakteristika promatrane pojave ili procesa. U kontekstu strojnog učenja tretira se kao nezavisna varijabla koja izravno utječe na predviđanja modela. U početku se pridobivanje značajki iz sirovih podataka za potrebe strojnog učenja obavljalo ručno od strane korisnika. Taj proces zove se proces inženjering značajki i uključuje nekoliko koraka od kojih su najvažniji ocjena značajki (engl. Feature evaluation), odabir značajki, rekombiniranje postojećih i izgradnja novih značajki. Učenje značajki (engl. Feature learning) slijedi te principe na automatiziran način te omogućuje proširenu reprezentaciju podataka i poboljšanu učinkovitost modela strojnog učenja [1]. Te metode zamjenjuju tradicionalne pristupe inženjeringa značajki u mnogim aplikacijama strojnoga učenja, od prepoznavanja govora i računalnog vida, do opće obrade signala [2]. Metode učenja značajki se mogu podijeliti na nenadzirane i nadzirane. Nenadzirani pristupi uče iz neoznačenih podataka, dok nadzirani pristupi uče iz označenih. Nadzirani pristupi omogućavaju

evaluaciju izrađenih modela koji se mogu izravno evaluirati u smislu pogreške u označavanju uzoraka učenja i time, posredno, ocijeniti kvalitete značajki. Ovaj rad fokusira se samo na nadzirane tehnike učenja značajki, specifično na problem klasifikacije. Posebna pozornost stavljena je na objašnjivosti tih metoda što je bitno za kredibilitet njihovih odluka. Objašnjivost se odnosi na stupanj do kojeg čovjek može razumjeti uzrok odluke koju je donio određeni model. Međutim, takvi modeli su uglavnom "crne kutije" (engl. Black-box) unatoč tome što su danas u širokoj uporabi. Kako bi se modelu vjerovalo važno je razumjeti zašto je model donio određene odluke. Istodobno, objašnjivost omogućuje i stjecanje strukturiranih i smislenih informacija i obrazaca iz podataka što vodi do potencijalnog otkrivanja novih znanja.

U postojećem stanju tehnike učenja značajki mogu se istaknuti sljedeći pristupi:

- odabir značajki (engl. Feature selection),
- nadzirno smanjenje dimenzionalnosti (engl. Supervised dimensionality reduction),
- nadzirno učenje rječnika (engl. Supervisd dictionary learning) i
- duboko učenje (engl. Deep learning).

2. PRISTUPI UČENJA ZNAČAJKI
2. FEATURE LEARNING APPROACHES

2.1. METODE ODABIRA ZNAČAJKI
2.1. FEATURE SELECTION

Metode odabira značajki smanjuju dimenzionalnost prostora ulaznih značajki odabirom podskupa važnih značajki i uklanjanjem nevažnih ili suvišnih [4]. To poboljšava računsku učinkovitost učenja klasifikacijskih modela. Uobičajeno se odabir značajki tretira kao problem optimizacije odabira podskupa značajki koje maksimiziraju učinkovitost predikcijskog modela. Metode odabira značajki dijele se u tri skupine: filtri, metode omotača i ugrađene metode. Te metode su objašnjive jer je iz ocjene individualne značajke ili podskupa značajki očit njihov utjecaj na klasifikaciju. Međutim, rezultati klasifikacije tih metoda su ograničeni jer ne mogu uvesti nove ili rekombinirati postojeće značajke.

2.2. NADZIRNO SMANJENJE DIMENZIONALNOSTI
2.2. SUPERVISED DIMENSIONALITY REDUCTION

Za razliku od metoda odabira značajki tehnike smanjenja dimenzionalnosti omogućuju rekombinaciju postojećih značajki. U posljednje vrijeme se broj značajki koje se mjere popeo s nekoliko desetaka na nekoliko stotina ili čak nekoliko tisuća. Povećanje dimenzija može, u teoriji, dodati više korisnih informacija klasifikatoru i time poboljšati njegovu točnost. Međutim, kod analize takvih podataka očekuje se rijetka reprezentacija za učenje što uzrokuje problem prokletstva dimenzionalnosti (engl. Curse of dimensionality) [3,4] te se zbog takvog nereprezentativnog uzorkovanja izvodi smanjivanje dimenzionalnosti. Osnovna zadaća tehnika nadziranog smanjenja dimenzionalnosti je pronaći učinkovito preslikavanje s kojim se izračunava niskodimenzionalna struktura podataka tako da se očuva što više korisnih informacija skrivenih u visokodimenzionalnom prostoru. Rezultat toga su nove značajke koje su linearne ili nelinearne rekombinacije ulaznih značajki [5].

Tipičan predstavnik linearnih metoda nadzirnoga smanjenja dimenzionalnosti je linearna diskriminantna analiza [6]. Metoda pronalazi linearne kombinacije značajki u niždimenzionalnom prostoru koje najbolje odvajaju uzorke određenih razreda. Preslikavanje visokodimenzionalnih uzorka u niže dimenzije temelji se na izračunu transformacijske matrice sastavljene iz svojstvenih vektora matrice međurazrednih (engl. Inter-class variance) i unutarrazrednih (engl. Intra-class variance) varijacija ulaznih značajki. Zbog linearnosti i pretpostavke o normalnoj distribuciji podataka ta je tehnika objašnjiva, ali su rezultati klasifikacije osjetljivi na korelirane značajke. Ta slabost je ispravljena [7] računanjem linearnih kombinacija značajki koje su visoko u korelaciji s oznakama razreda u niždimenzionalnom prostoru. Preslikavanje se izvede s izračunom transformacijske matrice unakrsne kovarijance (engl. Cross-covariance) između uzoraka značajki i oznaka razreda. Zbog upotrebe takve matrice pristup je neobjašnjiv u slučaju velikog broja

značajki, dok klasifikacijski modeli izgrađeni na rezultatima te metode često pate od pretjerane prilagodbe (engl. Overfitting). Pretjerana prilagodba je također slabost pristupa predloženog u članku [8] u slučaju malog broja uzoraka učenja iako je relativno jednostavan za interpretaciju. Metoda preslikava ulazne značajke u prostor nižih dimenzija tako da najbolje odvajaju uzorke pojedinih razreda, a da su u isto vrijeme takve linearne kombinacije ulaznih značajki visoko korelirane s oznakama razreda.

Do sad predstavljeni postupci ne mogu iskoristiti sve informacije skrivene u visokodimenzionalnom prostoru i stoga su ograničeni u klasifikacijskoj učinkovitosti. Značajan utjecaj na točnost klasifikacije proizlazi iz uspješnog identificiranja nelinearnih odnosa između značajki i oznaka razreda. Primjer takve metode predstavljen je u [9]. Metodom se pronalazi optimalnija geometrijska struktura ulaznih podataka koja zadržava izvorne udaljenosti između uzoraka za učenje pri preslikavanju u niže dimenzije, dok se udaljenost između uzoraka za učenje istog razreda smanjuje. To se postiže izgradnjom grafa susjedstva gdje čvorovi predstavljaju ulazne uzorke, dok su težine rubova između čvorova određene euklidskim udaljenostima između ulaznih uzoraka. Iz grafa susjedstva se izračuna matrica najkraćih puteva između svih parova čvorova u kojima je uključena i informacija o razredu. Matrica transformacije se zatim izračuna pronalaženjem svojstvenih vektora matrice najkraćeg puta. Budući da takav postupak zahtijeva velik broj izračunavanja udaljenosti i najkraćeg puta, metoda je vrlo spora u slučaju velikog broja uzoraka i značajki. Taj problem je riješen u [10], gdje autori predlažu učinkovitiju iterativnu metodu koja osim nelinearnih, razmatra i linearne odnose između susjednih uzoraka. Slično prethodnom postupku, i u ovom se slučaju izgradi graf susjedstva. Za preslikavanje u nižu dimenziju se, umjesto izračuna udaljenosti, izvede iterativni mehanizam rekonstrukcije pojedinačnih ulaznih uzoraka iz njegovih susjeda. Lokalna struktura susjedstva podataka zadržana je uključivanjem informacija o razredu. Unatoč učinkovitom izračunu ta metoda vrlo je osjetljiva na ulazne parametre, a posebice na veličinu susjedstva koja se koristi za rekonstrukciju. U slučaju velike vrijednosti tog parametra su kombinacije ulaznih značajki neobjašnjive. Metode nadzirnog smanjenja dimenzionalnosti su neobjašnjive i zbog toga jer preslikavanje u nižu dimenziju obično deformira udaljenosti između uzoraka učenja.

2.3. NADZIRNO UČENJE RJEČNIKA
2.3. SUPERVISED DICTIONARY LEARNING

U usporedbi s metodama predstavljenim u prethodnom poglavlju nadzirani algoritmi učenja rječnika omogućuju točniju klasifikaciju budući da mogu povećati ulazni prostor značajki. Ti pristupi uče rječnik i rijetku reprezentaciju (engl. Sparse representation) koji se koriste za predstavljanje ulaznih značajki kao linearne ili nelinearne kombinacije atoma rječnika. Atomi rječnika su sastavljeni iz svojstvenih vektora ulaznih podataka. Rijetke reprezentacije se iterativno ažuriraju tijekom procesa učenja kako bi se smanjila pogreška rekonstrukcije koju definiramo kao razliku između ulaznih uzoraka i rezultata rijetke reprezentacije [11].

Rječnik može biti "nedovoljno potpun" (engl. Undercomplete) pri čemu je dimenzija atoma manja od dimenzija ulaznih značajki. U slučaju korištenja nepotpunog rječnika radi se o smanjenju dimenzionalnosti. Nasuprot tome, može se izgraditi "prekompletan" rječnik (engl. Overcomplete) čije su dimenzije atoma veće od dimenzija ulaznih podataka. Upotreba prekompletnog rječnika proširuje informativnost ulaznog prostora što omogućuje veću učinkovitost klasifikacije. Ovisno o izboru mehanizma obrade diskriminatorne informacije, rječnici mogu biti zajednički (eng. Shared dictionary) ili razredno-specifični, odnosno izrađeni za svaki razred posebno (eng. Class specific dictionary). Korištenje zajedničkog rječnika zahtijeva uvođenje dodatnog klasifikatora dok razredno-specifični rječnik omogućuje izravnu klasifikaciju nepoznatih uzoraka na temelju pogreške rekonstrukcije [12].

Primjer metode izravne klasifikacije nepoznatih uzoraka na temelju pogreške rekonstrukcije predstavljen je u [13] gdje se razredno-specifični rječnici kreiraju zasebno nad svim uzorcima učenja. Razred nepoznatog uzorka određen je na temelju najniže pogreške rekonstrukcije koju predstave rječnici. Svaki razredno-specifični rječnik izgrađen je tako da optimalno

rekonstruira samo uzorke značajki određenog razreda. To može dovesti do povećanja broja izvedenih iteracija metode u slučaju velikog broja razreda što znatno povećava vrijeme učenja. Kako bi se ubrzao proces učenja, u [14] je predložena izgradnja razredno-specifičnih rječnika samo iz uzoraka učenja određenog razreda. Na taj način metoda preskače korak ažuriranja rječnika. Ovaj pristup je objašnjiv jer je iz pogreške rekonstrukcije izravno vidljivo kako određeni uzorak učenja utječe na klasifikaciju nepoznatog uzorka. Međutim, glavni nedostatak ovog pristupa je taj što ne može iskoristiti sve informacije iz ulaznih podataka, budući da uzorci značajki nisu potpuno neovisni između razreda i imaju zajedničke karakteristike [12].

Alternativna opcija je uvođenje pogreške klasifikacije kao dodatnog kriterija optimizacije tijekom učenja rječnika [15]. S ovim dodatnim kriterijem problem optimizacije postaje nekonveksan (engl. Non-convex) što znači da postoji više lokalnih optimuma koji nisu nužno globalni. Na primjer, u [16] je predstavljena metoda za klasificiranje rukom pisanih znamenki koja koristi učenje binarnih klasifikatora, po jedan za svaku znamenku (tj. razred), kao dodatni kriterij optimizacije. Glavni problem metode je što naizmjenično maksimizira rijetku reprezentaciju, minimizira pogrešku rekonstrukcije i uči klasifikator. Rezultat toga je velika količina parametara za podešavanje koji pridonose visokoj računskoj zahtjevnosti metode. Algoritam predstavljen u [17] pak istovremeno uči rječnik i parametre klasifikatora. To sprječava da metoda zapne pri učenju u lokalnom optimumu. U ovom su algoritmu upotrijebljeni samo linearni klasifikatori što uzrokuje loše klasifikacijske rezultate u slučajevima nelinearne klasifikacije. Rijetka reprezentacija također onemogućuje tumačenje dobivenog modela. Zbog toga je teško doći do korisnog znanja u onim slučajevima kada rječnik sadrži veliki broj atoma [12,13]

2.4. DUBOKO UČENJE
2.4. DEEP LEARNING

Svi do sada predstavljeni postupci koriste eksplicitno definirane linearne ili nelinearne postupke za postizanje rekombinacije značajki. Duboko učenje ima ugrađene mehanizme za dohvaćanje i linearnih i nelinearnih odnosa između značajki koji se postižu u umjetnim neuronima koji sastavljaju takve metode. U njima se ulazne značajke rekombiniraju množenjem i zbrajanjem težinskih vrijednosti uzoraka i njihovim normaliziranjem prema unaprijed definiranoj aktivacijskoj funkciji. Ovisno od aktivacijske funkcije, duboko učenje, odnosno neuronske mreže mogu obraditi nelinearne međuovisnosti između značajki [18].

Učenje neuronske mreže počinje nasumičnom inicijalizacijom težina veza između neurona i izvodi se iterativno podešavanjem istih. U svakoj iteraciji se izračunava razlika između željenog i stvarnog izlaza iz mreže uz pomoć funkcije gubitaka (engl. Loss function). Mehanizam učenja neuronskih mreža temelji se na minimizaciji te funkcije. Najraširenija strategija učenja neuronske mreže je algoritam povratnog širenja (engl. Backpropagation) [19]. Kod algoritma povratnog širenja se s metodom gradijentnog spusta (engl. Gradient descend) izračuna gradijent funkcije gubitka na temelju težina veza u svakoj iteraciji. S obzirom na gradijent funkcije gubitka se u koraku povratnog prolaza ažuriraju težine veze mreže [18] pomoću vanjskog parametra koji se naziva stopa učenja (engl. Learning rate). Algoritam završava kada se izvrši unaprijed određeni broj iteracija ili kada sustav dosegne stabilno stanje i težine mrežnih veza se više ne mijenjaju.

U [20] autori predlažu duboku neuronsku mrežu koja prilagođava težine veza za svaki uzorak učenja pomoću algoritma povratnog širenja. Iako to omogućuje konvergenciju parametara modela, takvo učenje zahtijeva veliki broj iteracija da bi se postigli optimalni rezultati. Autori u [21] stoga predlažu da se pojedinačna iteracija učenja izvede na većem broju uzoraka. Istodobno, predložena neuronska mreža automatski postavlja broj skrivenih slojeva i neurona u procesu učenja. Te razlike u arhitekturi pružaju bolji kompromis između vremena učenja i točnosti modela. Međutim, varijabilnost broja skrivenih slojeva i neurona doprinosi neobjašnjivosti naučenih značajki. Neobjašnjivost je također slabost pristupa predloženog u [22]. Prikazana duboka mreža sastoji se iz tri skrivena sloja neurona. Prvi

skriveni sloj proširuje dimenzionalnost ulaznog prostora, a drugi skriveni sloj grupira uzorke iz prethodnog sloja u klastere na temelju njihove međusobne sličnosti. Iz tih klastera se u sljedećem sloju agregiraju nove značajke koje učinkovitije diferenciraju uzorke određenih klasa. U slučaju takvog učenja postoji opasnost da se unutar istih klastera pojave uzorci iz različitih razreda što onemogućuje objašnjivanje naučenih značajki. U [22] je predloženi postupak izvodi učenja značajki s neuronskom mrežom kako bi se poboljšao interpretabilni model klasifikacijske metode najbližeg susjeda. Metoda koristi evolucijsko računanje za optimizaciju težina veze neuronske mreže, dok se izvedba optimizacije veza između neurona temelji na distribuciji najbližih susjeda. No ni u ovom slučaju autori ne daju objašnjenja naučenog modela.

3. PRISTUPI OBJAŠNJIVANJA
3. TOOLS FOR EXPLAINABILITY

Objašnjenje odluka ranih pristupa strojnog učenja prilično je jednostavna. Stabla odlučivanja (engl. Decision trees) je moguće analizirati neposrednim prijelazom njihovih čvorova uz izračun uobičajene lokalne i globalne klasifikaciju greške [24, 25]. Linearni modeli se objasne s pogledom u njihove parametre [26]. Unatoč tome, takve vrste modela ne izvlače u dovoljnoj mjeri svo znanje koje sadrži podatke i često postižu niske klasifikacijske točnosti. Metode koje su bolje u dobivanju znanja iz podataka postižu i više točnosti, ali su obično „crne kutije“ jer pored složenih linearnih odnosa razmatraju i nelinearnosti u podacima [24]. Zato je nastala potreba za uvođenje pristupa za interpretaciju takvih modela. To se postiže ili učenjem interpretativnih modela oko danog uzorka interesa [27] ili s ocjenom važnosti svake pojedinačne značajke pri klasifikacijskom izlazu [28]. Ocjena važnosti svake pojedinačne značajke temelji se na Shapleyjevim vrijenostima koje prikazuju kako individualna značajka doprinosi odluci modela s opažanjem razlika među vjerojatnosti očekivane i stvarne klasifikacije kada promatranu značajku otklonimo. Obje tehnike koriste se za interpretaciju ulaznih značajki i osjetljive su na korelirane značajke što onemogućuje njihovu primjernu za interpretaciju naučenih značajki.

Pristup predstavljen u [29] uči nekorelirane visokokvalitetne značajke iterativno na slijedni način. Te naučene značajke predstavljene su u formi jednadžba kao linearne i nelinearne kombinacije ulaznih. To omogućava korištenje obje metode za interpretaciju naučenih značajki. Neuronske mreže se mogu objasniti vizualizacijom njihovog procesa odlučivanja. Obično se ostvaruje stvaranjem toplotnih karata (angl. Heat map) iz ulaznih uzoraka, s naglaskom na uzorke koji maksimiziraju aktivaciju promatranih neurona. Pokazalo se da su te tople karte kontra intuitivne [30], a sam pristup je ograničen samo na upotrebu kod neuronskih mreža [31]. U zadnjem vremenu predlažu se različite interpretabilne arhitekture neuronskih mreža, koje identificiraju značajke s visokom intraklasne varijancom na ulaznim podacima i rekombiniraju ih s značajkama niske interklasne varijance kako bi se poboljšalo razumijevanje predviđanja [32]. Iako to omogućuje proučavanje načela odlučivanja za neuronske mreže i utjecaj ulaznih značajki na klasifikaciju, takve naučne značajke su još uvijek vrlo apstraktne i nisu primjerene za otkrivanje znanja.

4. ZAKLJUČAK
4. CONCLUSION

Postojeći pristupi učenja značajki su ili neobjašnjivi ili ograničeni u svojoj točnosti to je posljedica nemogućnosti rekombiniranja ulaznih značajki. S druge strane, pristupi koji omogućuju povećanje dimenzionalnosti prostora ulaznih značajki su obično računalno zahtjevni, budući da zahtijevaju iterativne nekonveksne optimizacije i podešavanje mnogih konfiguracija skrivenih dimenzija. Za takve pristupe postoje alati kojima se u određenoj mjeri mogu objasniti razlozi odluka prediktivnih modela. Te tehnike rješavanju problem interpretacije modela korištenjem alata za izračun važnosti svake pojedinačne značajke na klasifikacijskom izlazu ili s učenjem interpretativnih modela oko uzoraka interesa. Iako su te metode ograničene samo na ulazne značajke, s njihovom pravilnom upotrebom moguće ih je primijeniti i na naučene značajke. Tehnike za objašnjenje naučenih

značajki kod neuronskih mreža se ostvaruje stvaranjem toplinskih karata koji maksimiziraju aktivaciju promatranih neurona. Svejedno su u tom primjeru objašnjenja vrlo apstraktna i nisu primjerena za otkrivanje znanja.

5. REFERENCE

5. REFERENCES

[1.] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798-1828, 2013., DOI: 10.1109/TPAMI.2013.50

[2.] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 2019., DOI: https://doi.org/10.3390/electronics8080832

[3.] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics), page 728. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2006., ISBN 978-0-387-31073-2

[4.] Huan Liu and Hiroshi Motoda. Computational Methods of Feature Selection, page 440. Chapman & Hall/CRC, 1st edition, 2007., ISBN-13: 978-0387-31073-2

[5.] Yunqian Ma and Yun Fu. Manifold Learning Theory and Applications, page 314. CRC Press, Inc., USA, 1st edition, 2011., ISBN: 9781439871096

[6.] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. Linear discriminant analysis: A detailed tutorial. AI Communications, 30(2):169-190, 2017., DOI: 10.3233/AIC-170729

[7.] Marco Loog, Bram van Ginneken, and Robert P. W. Duin. Dimensionality reduction by canonical contextual correlation projections. In Tomás Pajdla and Jirí Matas, editors, Computer Vision - ECCV 2004, pages 562-573, 2004., DOI: 10.1007/978-3-540-24670-1_43

[8.] Richard G. Brereton and Gavin R. Lloyd. Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics, 28(4):213-225, 2014., DOI: 10.1002/cem.2609

[9.] Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou. Supervised nonlinear dimensionality reduction for visualization and classication. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) , 35(6):1098-1107, 2005., DOI: 10.1109/TSMCB.2005.850151

[10.] Yuanhong Liu, Yansheng Zhang, Zhiwei Yu, and Ming Zeng. Incremental supervised locally linear embedding for machinery fault diagnosis. Engineering Applications of Artificial Intelligence, 50:60-70, 2016., DOI: 10.1016/j.engappai.2015.12.010

[11.] Mehrdad J. Gangeh, Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. Supervised dictionary learning and sparse representation-a review. ArXiv, abs/1502.05928, 2015.

[12.] W. Tang, A. Panahi, H. Krim, and L. Dai. Analysis dictionary learning based classification: Structure for robustness. IEEE Transactions on Image Processing, 28(12):60356046, 2019., DOI: 10.1109/TIP.2019.2919409

[13.] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Fisher discrimination dictionary learning for sparse representation. In Proceedings of the IEEE International Conference on Computer Vision, pages 543-550, 2011., DOI: 10.1109/ICCV.2011.6126286

[14.] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210-227, 2009., DOI: 10.1109/TPAMI.2008.79

[15.] Mehrdad J. Gangeh, Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. Supervised dictionary learning and sparse representation-a review. ArXiv, abs/1502.05928, 2015.

[16.] Haoli Zhao, Peng Zhong, Haiqin Chen, Zhenni Li, Wuhui Chen, and Zibin Zheng. Group non-convex sparsity regularized partially shared dictionary learning for multi-view learning. Knowledge-Based Systems, 242(C):1-16, 2022., DOI: 10.1016/j.knosys.2022.108364

[17.] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R. Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Proceedings of the 21st International Conference on Neural Information Processing Systems, pages 1033-1040, 2009., DOI: https://doi.org/10.48550/arXiv.0809.3083

[18.] Michael A. Nielsen. Neural Networks and Deep Learning, page 216. Determination Press, 2018.

[19.] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323:533-536, 1986., DOI: https://doi.org/10.1038/323533a0

[20.] Luis Miralles-Pechuán, Dafne Rosso, Fernando Jiménez, and Jose M. García. A methodology based on Deep Learning for advert value calculation in CPM, CPC and CPA networks. Soft Computing, 21(3):651-665, 2017., DOI: 10.1007/s00500-016-2468-4

[21.] S. Jaiyen, C. Lursinsap, and S. Phimoltares. A very fast neural learning for classification using only new incoming datum. IEEE Transactions on Neural Networks, 21(3):381-392, 2010., DOI: 10.1109/TNN.2009.2037148

[22.] Hung-Wen Peng, Shie-Jue Lee, and Chie-Hong Lee. An oblique elliptical basis function network approach for supervised learning applications. Applied Soft Computing, 60:552-563, 2017., DOI: 10.1016/j.asoc.2017.07.019

[23.] Lin Wang, Bo Yang, Yuehui Chen, Xiaoqian Zhang, and Jeff Orchard. Improving Neural-Network Classifiers Using Nearest Neighbor Partitioning. IEEE Transactions on Neural Networks and Learning Systems, 28(10):2255-2267, 2017., DOI: 10.1109/TNNLS.2016.2580570

[24.] Charu C. Aggarwal. Data Mining: The Textbook, page 734. Heidelberg: Springer, 2015.

[25.] Mohammad Azad, Igor Chikalov, and Mikhail Moshkov. Representation of knowledge by decision trees for decision tables with multiple decisions. Procedia Computer Science, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. 176:653-659, 2020., DOI: https://doi.org/10.1016/j.procs.2020.09.037

[26.] R. Chandler. On the use of generalized linear models for interpreting climate variability. Environmetrics, 16(7):699-715, 2005., DOI: 10.1002/env.731

[27.] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 4768-4777, 2017., DOI: https://doi.org/10.48550/arXiv.1705.07874

[28.] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41:647-665, 2013., DOI: DOI: 10.1007/s10115-013-0679-x

[29.] D. Vlahek and D. Mongus,. An Efficient Iterative Approach to Explainable Feature Learning. IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 5, pp. 2606-2618, 2023., DOI: 10.1109/TNNLS.2021.3107049

[30.] Mathieu Aubry and Bryan Russell. Understanding deep features with computer-generated imagery. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 2875-2883, 2015., DOI: https://doi.org/10.48550/arXiv.1506.01151

[31.] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5188-5196, 2015., DOI: https://doi.org/10.48550/arXiv.1412.0035

[32.] Dahuin Jung, Jonghyun Lee, Jihun Yi, and Sungroh Yoon. icaps: An interpretable classifier via disentangled capsule networks. In Computer Vision - ECCV 2020 - 16th European Conference, volume 12364 of Lecture Notes in Computer Science, pages 314-330, 2020., DOI: https://doi.org/10.48550/arXiv.2008.08756

AUTORI · *AUTHORS*

• **Dino Vlahek** - viši istraživač u Laboratoriju za geoprostorno modeliranje, multimediju i umjetnu inteligenciju na Fakultetu za elektrotehniku, računarstvo i informatiku Sveučilišta u Mariboru. Njegovi istraživački interesi su učenje značajki, analitika podataka i interpretacija modela. Magistrirao je 2018. na Fakultetu za elektrotehniku, računarstvo i informatiku Sveučilišta u Mariboru, dok je na istom fakultetu doktorirao 2024. godine, smjer Računarstvo i informatika.

Korespondencija · *Correspondence*

dino.vlahek1@um.si

• **Bojan Nožica** - nepromijenjena biografija nalazi se u časopisu Polytechnic & Design Vol. 7, No. 1, 2019.

Korespondencija · *Correspondence*

bojan.nozica@tvz.hr