Inženjering karakteristika: Implikacije selekcije karakteristika na performanse predikcije

Olivera Janković
ORAO a.d.
Bijeljina, RS, BiH
olivera.jankovic@orao.aero

Đorđe Babić
RAF
Beograd, Srbija
djbabic@raf.rs

Sažetak— **Potreba da se iz raspoloživih skupova podatka dobiju karakteristike koje će omogućiti bolje prediktivne performanse sa jedne, i manje složene i efikasne modele sa druge strane, sastavni je dio kompromisa i stalni izazov procesa prediktivnog modelovanja. U radu će biti prikazan odabir karakteristika, korištenjem tehnike nenadziranog učenja - analize glavnih komponenti (opcija izlaza podskup karakteristika koji nude većinu informacija), te implikacije redukcije dimenzionalnosti na performanse predikcije kvara avionskog motora, podacima vođenog modela održavanja.**

Ključne riječi - inženjering karakteristika; selekcija karakteristika; analiza glavnih komponenti PCA;

I. UVOD

Performanse metoda mašinskog učenja u velikoj mjeri zavise od izbora predstavljanja (*representation*) podataka (ili karakteristika) na koje se one primjenjuju. Iz tog razloga, veliki dio stvarnih napora u primjeni algoritama mašinskog učenja odlazi u dizajn procesa preprocesiranja i transformacija podataka, koji rezultiraju prikazom podataka koji mogu da podrže efektivno mašinsko učenje. Inženjering karakteristika predstavlja način da se iskoristi ljudski um, domišljatost i prethodno znanje kako bi se nadoknadila ta slabost. U osnovi zahtijeva i znanje o tome kako funkcioniše algoritam mašinskog učenja, jer različiti algoritmi zahtjevaju različite metode inženjeringa karakteristika. U samom procesu obuke modela potrebni su efikasni algoritmi za rješavanje problema optimizacije, kao i za skladištenje i procesiranje masivne količine raspoloživih podataka. Reprezentacija i algoritamsko rješenje za zaključivanje obučenog modela moraju biti efikasni; u određenim aplikacijama, efikasnost algoritma učenja ili zaključivanja, odnosno njegova prostorna i vremenska kompleksnost može biti toliko važna kao i njegova preciznost [1].

Inženjering karakteristika nije formalno definisan termin; u osnovi to je mnogo zadataka vezanih za dizajniranje skupova karakteristika za aplikacije mašinskog učenja. Generalno, potrebno je mnogo vremena i truda u vezi sa inženjeringom karakteristika, koji često zahtjeva vještine baze podataka, kodiranja i programiranja; postoji mnogo pokušaja i grešaka za testiranje uticaja novokreiranih karakteristika. To može biti repetitivan posao, nakon koga se može doći do spoznaje da dodavanje više karakteristika može da vodi i ka pogoršanju tačnosti.

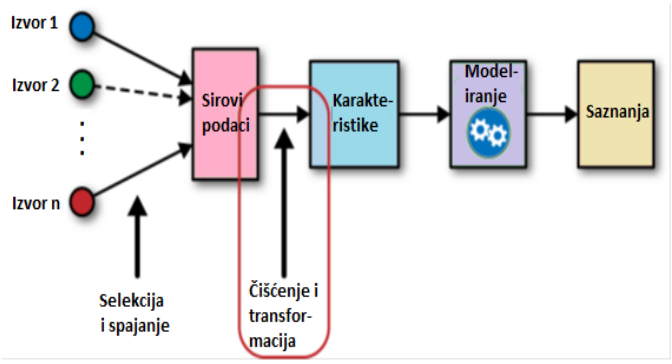
Odabir relevantnih karakteristika i uklanjanje irelevantnih je važan problem mašinskog učenja, odnosno važan korak u izgradnji prediktivnih modela; praktični algoritmi mašinskog učenja često mogu da smanje performanse kada su suočeni sa mnogim karakteristikama, koje nisu potrebne za predikciju željenog rezultata. Stoga je potrebno primjeniti postupak uklanjanja ovakvih karakteristika kako bi se povećala efikasnost, poboljšala preciznost i spriječilo prekomjerno podudaranje (*overfitting*). U radu će biti prikazan način selekcije karakteristika, korištenjem tehnike nenadziranog učenja - analize glavnih komponenti (koristi se opcija izlaza podskupa karakteristika koji nude većinu informacija), smanjenje dimenzionalnosti i implikacije na performanse predikcije za potrebe predikcije kvara avionskog motora (u skladu sa prirodom seta podataka za trening), sa ciljem izbora podskupa originalnih karakteristika skupa podataka i podacima vođenog (*data driven*) modela održavanja [2], tako da nad njima primjenjeni algoritmi mašinskog učenja, generišu klasifikator kojim može da se minimizira kompromis (*trade off*) u složenosti i tačnosti.

II. INŽENJERING KARAKTERISTIKA

Da bi se razumio sam termin inženjering karakteristika, kako bi se postigla šira slika aplikacije, korisno je kratko sagledati sistem mašinskog učenja, njegove osnovne pojmove kao što su podaci i modeli. Mašinsko učenje u osnovi spreže matematičke modele i podatke, kako bi se dobila određena saznanja ili napravila predikcija. Kao ulaz ovi modeli uzimaju karakteristike (*feature*), koje predstavljaju numeričku reprezentaciju jednog od aspekata, neobrađenih, sirovih podataka [3], koji su u osnovi opažanja pojava stvarnog svijeta; svaki podatak predstavlja mali prozor ograničenog aspekta stvarnosti, zbirka svih tih opažanja daje u osnovi nejasnu sliku cjeline, sastavljenu od velikog broja dijelova, praćenu pojavom šuma mjerenja i dijelova koji nedostaju. U tom procesu veoma je važno znati i razlog, odgovor na pitanje, zašto se podaci prikupljaju. Sam put od podataka do odgovora pun je obećavajućih početaka i slijepih ulica; radni tokovi sa podacima često su višeetapni, iterativni procesi. Pri tome, treba se nositi sa kvalifikacijama podataka, kao što su: pogrešni

(*wrong*) - podaci su rezultat greške u mjerenju, redundantni (*redundant*) podaci - sadrže više aspekata koje pružaju potpuno istu informaciju ili nedostajući (*missing*) podaci – ukoliko informacije nisu prisutne za neke tačke podataka.

Obzirom da sirovi podaci (*raw data*) često nisu numerički, uvode se karakteristike kao numerički prikaz sirovih podataka. Postoje mnogi načini da se neobrađeni, sirovi podaci pretvore u numerička mjerenja, zbog čega karakteristike mogu imati različite oblike, pri čemu moraju proizilaziti iz vrste dostupnih podataka. Iako je, možda, manje očita, stoji činjenica da su karakteristike povezane sa modelom; neki su modeli prikladniji za neke tipove karakteristika i obrnuto. Inženjering karakteristika je proces formulisanja najprikladnijih karakteristika s obzirom na podatke, model i zadatak. Važan je, takođe, broj karakteristika - ako nema dovoljno informativnih karakteristika, onda model neće biti u stanju da izvrši postavljeni zadatak, ako ima previše karakteristika ili ako je većina njih irelevantna, model će biti skuplji i zahtjevniji za obuku. Na Sl.1 može se vidjeti da se karakteristike i modeli nalaze između sirovih podataka i željenih spoznaja, odnosno inženjering karakteristika se nalazi između podataka i modeliranja u okviru procesa mašinskog učenja za dobivanje saznanja (*insights*). U toku procesa mašinskog učenja odabire se ne samo model, već i karakteristike; izbor jednog utiče na drugi. Dobre karakteristike čine sljedeći korak modeliranja jednostavnijim, a rezultirajući model sposobnijim ispuniti željeni zadatak. Loše karakteristike mogu zahtjevati mnogo složeniji model da bi se postigao isti nivo performansi. Nadalje, autori u knjizi [3], obrađuju različite vrste karakteristika i diskutuju njihove prednosti i mane za različite vrste podataka i modela.



Slika 1. Mjesto inženjeringa karakteristika u procesu mašinskog učenja [2]

Nargesian et al. [4] navode da je inženjering karakteristika centralni zadatak u pripremi podataka za mašinsko učenje; praksa je da se konstruišu prikladne, podesne karakteristike iz zadatih karakteristika koje dovode do poboljšanih prediktivnih performansi. Inženjering karakteristika tako uključuje primjenu funkcija transformacije (npr. aritmetički i agregatni operatori), koje pomažu u skaliranju karakteristike ili pretvaranju nelinearne relacije između karakteristike i ciljne klase u linearni odnos, koji je lakše naučiti. Autori navode da je inženjering karakteristika zadatak poboljšanja performansi prediktivnog modeliranja na skupu podataka, transformisanjem njegovog prostora karakteristika (*feature space*). Autori predstavljaju tehniku, nazvanu učenje inženjeringa karakteristika (LFE, Learning Feature Engineering), za automatizaciju inženjeringa karakteristika zadataka klasifikacije. Za izvođenje automatizovanog inženjeringa karakteristika, neki postojeći pristupi usvajaju vođeno pretraživanje u prostoru karakteristika koristeći heurističke mjere kvaliteta karakteristika, kao što je informacioni dobitak i druge zamjenske mjere performansi [5]. Khurana et al. [6] izvode pohlepnu konstrukciju i odabir karakteristika na osnovu evaluacije modela i predstavljaju novi sistem "Cognito", koji obavlja automatski inženjering karakteristika na datom skupu podataka za nadzirano učenje. U [7] su razvili mašinu naučnih podataka (DSM, Data Science Machine), koja je u stanju da automatski izvede prediktivne modele iz sirovih podataka, koja razmatra problem inženjeringa karakteristika kao izbor karakteristika na prostoru novih karakterisika. Duboke neuronske mreže (DNN, Deep Neural Networks) automatski uče korisne karakteristike [8] i pokazale su zapažene uspjehe na video, slikovnim i govornim podacima; međutim, u nekim domenima je još uvijek potreban inženjering karakteristika (npr. domen zdravstva). U slučaju probabilističkih modela, dobra reprezentacija je često ona koji obuhvata posteriornu (naknadnu) distribuciju osnovnih faktora objašnjenja za posmatrani ulaz; dobra reprezentacija je korisna i kao ulaz za nadzirani (*supervised*) prediktor.

Za većinu tipova otkrivanja kvarova i aplikacija za prediktivno održavanje, vrijednosti dobijene iz sistema za prikupljanje podataka (*data acquisition system*) obično moraju biti prethodno obrađene prije njihovog transformisanja u novi prostor varijabli, za bolje performanse algoritma mašinskog učenja. Najvažniji procesi inženjeringa karakteristika u kontekstu detekcije tipa kvara i prediktivnog održavanja, bazne namjene korištenog seta podataka (podacima vođen model predikcije kvara) ovoga rada, u osnovi su: Obrada signala (*Signal Processing*) - Interpretacija, generisanje i transformacija sirovih neobrađenih podataka; Ekstrakcija karakteristika (*Feature Extraction*) – Generisanje novih informacije kombinovanjem karakteristika i Selekcija karakteristika (*Feature Selection*) – selekcija, izbor podskupa najreprezentativnijih karakteristika.

III. SELEKCIJA KARAKTERISTIKA

Rad [9] objavljen 1997. godine, koji predstavlja posebno izdanje o relevantnosti karakteristika, koristio je nešto više od 40 karakteristika; rad sadrži članke sa definicijama i diskusijama različitih pojmova relevantnosti. Situacija se znatno promjenila sa godinama. Izbor varijabli i karakteristika postao je fokus mnogih istraživanja u oblastima primjene za koje su dostupni podaci sa desetinama ili stotinama hiljada promjenljivih. Ta područja uključuju obradu teksta na internetskim dokumentima, analizu ekspresije gena itd. U osnovi cilj izbora varijabli/karakteristika je trostruki: poboljšanje performansi predviđanja (predikcionih performansi) prediktora, obezbjeđivanje bržih i ekonomičnijih prediktora, i obezbeđivanje boljeg razumjevanja osnovnog/temeljnog procesa koji je generisao podatke. Rad [9] pokriva širok spektar aspekata takvih problema: obezbjeđivanje bolje definicije ciljne funkcije, konstrukciju karakteristika (*feature construction*), rangiranje karakteristika (*feature ranking*), izbor višestrukih (*multivariate*)

karakteristika, efikasne metode pretraživanja i metode procjene validnosti karakteristika.

Postoje mnoge potencijalne prednosti selekcije varijabli i karakteristika: olakšavanje vizuelizacije podataka i razumjevanja podataka, smanjenje zahtjeva za mjerenjem i skladištenjem, smanjenje vremena obuke/treninga i korišćenja, prkoseći problemu dimenzionalnosti radi poboljšanja performansi predikcije. Neke metode stavljaju veći naglasak na jedan aspekt od drugog; čest je fokus na konstruisanje i odabir podskupova karakteristika koje su korisne za izgradnju dobrog prediktora.

Mnogi algoritmi odabira varijabli uključuju rangiranje varijabli kao glavni ili pomoćni mehanizam odabira zbog svoje jednostavnosti, skalabilnosti i dobrog empirijskog uspjeha; Razmatra se skup *m* primjera $\{x_k, y_k\}$ $(k = 1,...m)$ koji se sastoji od *n* ulaznih varijabli $x_{k,i}$ $(i = 1,...n)$ i jedne izlazne varijable y_k. Rangiranje varijabli koristi funkciju bodovanja *S(i)* izračunatu iz vrijednosti $x_{k,i}$ i y_k, $k = 1,...m$. Prema konvenciji, pretpostavlja se da je visoka ocjena indikativna za vrijednu (*valuable*) varijablu i da se varijable sortiraju u padajućem redu *S(i)*. Da bi se koristilo rangiranje varijabli za izgradnju prediktora, definisani su ugnježđeni podskupovi koji progresivno uključuju sve više i više varijabli smanjene relevantnosti.[10]

Postoje dvije glavne grupe mehanizama redukcije: dobijanje atributa, podgrupa koje su adekvatnije za predviđanje i transformacija iz skupa podataka u niži dimenzioni prostor. Prvo je smanjenje kandidata nezavisnih varijabli u odnosu na zavisnu varijablu. Nadgledane (Supervised) tehnike su filter [11] i wrapper [12]. Ove metode selekcije karakteristika odabiru skup karakteristika iz postojećih karakteristika i ne grade nove (nema ekstrakcije karakteristika).Drugo, transformacija skupa podataka u niži dimenzioni prostor, je smanjenje skupa varijabli u manji skup uz zadržavanje većine informacionog sadržaja. Kako bi se slijedile ove nenadzirane (Unsupervised) tehnike za smanjenje dimenzija, postoje uglavnom dvije metode, a to su faktorska analiza (izlaz faktori) i analiza glavnih komponenti (daje komponente).

A. Analiza glavnih komponenti

Analiza glavnih komponenti (PCA, Principal Component Analysis) je tehnika izdvajanja karakteristika koja generiše nove karakteristike koje su linearna kombinacija početnih karakteristika. PCA mapira svaku instancu datog skupa podataka prisutnu u d dimenzionalnom prostoru na k dimenzionalni podprostor tako da je $k < d$. Generisan je skup k novih dimenzija koje se nazivaju glavne komponente (PC), a svaka glavna komponenta je usmjerena ka maksimalnoj varijansi, izuzev varijanse koja je već uzeta u sve prethodne komponente. Nakon toga, prva komponenta pokriva maksimalnu varijansu, a svaka komponenta koja slijedi pokriva manju vrijednost varijanse.

PCA je široko korištena multivarijantna analitička statistička tehnika koja se može primijeniti za smanjenje skupa zavisnih varijabli na manji skup osnovnih varijabli (zvanih komponente) na osnovu uzoraka korelacije između izvornih varijabli.[13]

IV. EKSPERIMENTALNE POSTAVKE I REZULTATI

Osnovne karakteristike korištenog seta podataka, koji predstavlja vremenske serije turbogasnih motora aviona, nastale simulacijom, su:
Izvor/ripozitorijum podataka: NASA Ames (2017.).
Sirovi podaci:

- Train_FD003.txt - za trening, sadrži 24.720 instanci i 26 atributa;
- Test_FD003.txt - za testiranje, sadrži 16.596 instanci i 26 atributa

Preprocesiranje podataka:

- Proces inženjeringa karakteristika – agregacija, srednja vrijednost i standardna devijacija, vremenski prozor W=10.

Označavanje podataka:

- Kreiranje labela - kontekst predviđanja kvara unutar 30 ciklusa.

Na ulazni skup podataka, primjenjen je algoritam Analiza glavnih komponenti (R komande u Weka konzoli). Postoji opcija da izlaz bude linearna transformacija ulaznih karakteristika odnosno izlaz analize glavnih komponenti je skup komponenti formiranih pomoću linearnih kombinacija koreliranih atributa. Obzirom na to da postoji set test podataka koji će se koristiti za evaluaciju klasifikatora (ne koristi se unakrsna validacija za evaluaciju klasifikatora) odabrana je simulacija, u kojoj je izlaz podskup karakteristika, koji nude većinu informacija. Za odabranu vrijednost varijanse (*variance*) 95% dobijen je podskup od 54 karakteristike (55 računajući atribut klase), odnosno, smanjena je dimenzionalnost ulaznog skupa sa 69 na 55 karakteristika. Pored toga, rangirane su karakteristike podskupa podataka u skladu sa vrijednošću varijanse od 95%, što je prikazano u Tabeli I.

Prvotno smanjenje dimenzionalnosti, (54 karakteristike, Train_FD003_54, evaluacija korištenjem Test_FD003_54) kao što se može vidjeti u Tabeli II, različito se odrazilo na korištene klasifikatore. Algoritam vektora podrške SMO je postigao manju tačnost 99.14% od 99.18%; redukcija dimenzije ulaznog skupa, korištenjem odabranog PCA izlaza, nije se odrazila na klasifikacionu tačnost višeslojnog perceptrona - MLP i algoritam Logistička regresija koji su postigli istu klasifikacionu tačnost od 99.03%. i 99.11% respektivno.

Kako bi se ispitala mogućnost generisanja klasifikatora, sa ciljem minimiziranja kompromisa u složenosti i tačnosti, nadalje je formirano, pored inicijalnih Train_FD003_54 i Test_FD003_54 koji sadrže po 54 karakteristike, dodatnih 10 trening i analogno njima 10 testnih skupova, koji sadrže: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 karakteristika, u skladu sa rangiranjem datim u Tabeli I. Eksperimentalni rezultati postignuti za date ulazne trening skupove Train_FD003_5 do Train_FD003_50, evaluacijom korištenjem testnih skupova Test_FD003_5 do Test_FD003_50 respektivno, primjenom korištenih klasifikatora (Logistička regresija, SMO (Weka implementacija SVM) i MLP) dati su u Tabeli II. Uvidom u iste generalno se može zaključiti da se proces odabira karakteristika:

- različito odrazio na performanse za različite prediktore;

TABELA I. PCA RANGIRANJE KARAKTERISTIKA

Rang	Karakteristika	Rang	Karakteristika
1.	SSDenzor21	28.	SSenzor9
2.	Senzor21	29.	SSenzor10
3.	Senzor17	30.	SSDenzor12
4.	Senzor15	31.	SSDenzor10
5.	Senzor14	32.	SSDenzor9
6.	Senzor20	33.	SSDenzor8
7.	SSenzor2	34.	SSDenzor11
8.	Senzor12	35.	SSDenzor13
9.	SSenzor3	36.	SSenzor11
10.	SSenzor7	37.	SSDenzor14
11.	SSenzor6	38.	SSDenzor17
12.	SSenzor4	39.	SSDenzor16
13.	Senzor13	40.	SSDenzor15
14.	Senzor11	41.	SSDenzor7
15.	SSDenzor20	42.	SSDenzor6
16.	Senzor3	43.	SSDenzor5
17.	Setovanje2	44.	SSDenzor4
18.	Setovanje1	45.	SSenzor14
19.	Ciklus	46.	SSenzor13
20.	Senzor2	47.	SSenzor12
21.	Senzor4	48.	SSenzor15
22.	Senzor10	49.	SSenzor17
23.	Senzor6	50.	SSenzor20
24.	Senzor9	51.	SSDenzor3
25.	Senzor8	52.	SSDenzor2
26.	Senzor7	53.	SSenzor21
27.	SSenzor8	54.	ID_motora

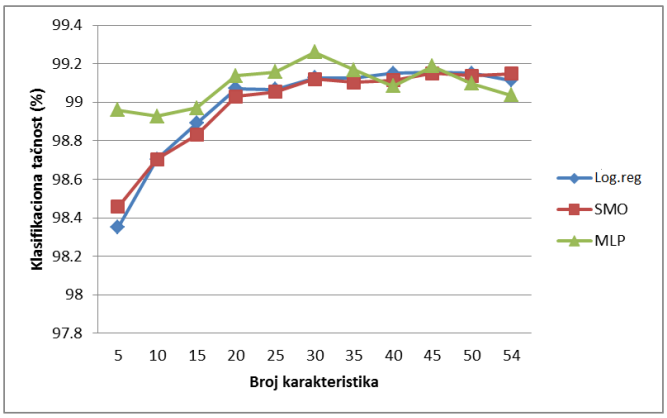
- postoje redukovani setovi podataka, koji sadrže takve karakteristike, korištenjem kojih se za svaki korišteni algoritam, generiše klasifikator kojim može da se minimizira kompromis u složenosti tačnosti (Sl. 2).

Uvidom u Tabelu II i sa Sl. 2 se, pored ostalog, može vidjeti da se za npr. višeslojni perceptron može odabrati, znatno manje složen model (u pitanju je algoritam, koji je od tri korištena vremenski najzahtjevniji i kojem se značajno povećava vrijeme izvršenja sa povećanjem broja karakteristika) već sa 20 karakteristika; set od 30 karakteristika, korištenjem MLP, postigao je najveću klasifikacionu tačnost od 99.25%.

TABELA II. REZULTATI KLASIFIKATORA (KVALIFIKACIONA TAČNOST) ZA SETOVE PODATKA NASTALE SELEKCIJOM KARAKTERISTIKA

Trening set	Testni set	Logistička regresija Klas. tačnost (%)	SMO Klas. tačnost (%)	MLP Klas. tačnost (%)
Train_FD003_5	Test_FD003_5	98.35	98.46	98.60
Train_FD003_10	Test_FD003_10	98.70	98.70	98.93
Train_FD003_15	Test_FD003_15	98.89	98.83	98.97
Train_FD003_20	Test_FD003_20	99.07	99.03	99.14
Train_FD003_25	Test_FD003_25	99.06	99.05	99.16
Train_FD003_30	Test_FD003_30	99.13	99.12	**99.26**
Train_FD003_35	Test_FD003_35	99.13	99.10	99.17
Train_FD003_40	Test_FD003_40	99.15	99.11	99.08
Train_FD003_45	Test_FD003_45	99.16	99.15	99.19
Train_FD003_50	Test_FD003_50	99.15	99.14	99.10
Train_FD003_54	Test_FD003_54	99.11	99.15	99.04
Train_FD003[a]	Train_FD003[b]	99.11	99.18	99.04

a.i b. Setovi podataka prije procesa selekcije karakteristika (69 karakteristika)



Slika 2. Grafički prikaz uticaja selekcije karakteristika za Log. reg., SMO i MLP klasifikatore (rangiranje karakteristika korištenjem PCA)

V. ZAKLJUČAK

Inženjering karakteristika, u osnovi, zahtijeva resurse, vještine nauke o podacima i znatnu vremensku posvećenost; primjena pokušaja i pogrešaka može potrajati i mjesecima. Uprkos važnosti, tema se rjetko razmatra sama, obzirom da je, zbog raznovrsnosti podataka i modela, problem generalizovati praksu inženjeringa karakteristika u svim projektima. Prave karakteristike je potrebno definisati u kontekstu i modela i podataka; u radu su prikazane smjernice za jedan segment procesa inženjeringa karakteristika, selekcija karakteristika u kontekstu predikcionog modela kvara avionskog motora, kojima se generiše klasifikator kojim može da se minimizira kompromis složenosti i tačnosti.

LITERATURA

[1] E. Alpaydın, Introduction to Machine Learning, 2nd ed., The MIT Press Cambridge, England, 2010. https://mitpress.mit.edu/contributors/ethem-alpaydin

[2] O. Janković, Đ. Babić, "Održavanje u avio-industriji i podacima vođeni modeli, " 4th International Scientific Conference COMETa 2018, pp 688-695, 2018. http://cometa.ues.rs.ba/Zbornik%20COMETa2018.pdf

[3] A. Zheng and A. Casari, Mastering Feature Engineering, 1st ed., Published by O'Reilly Media, 2018, https://perso.limsi.fr/annlor/enseignement/ensiie/Feature_Engineering_for_Machine_Learning.pdf

[4] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil and D. Turaga, "Learning Feature Engineering for Classification," IJCAI-17, pp 2529-2535, 2017. https://www.researchgate.net/publication/318829821_Learning_Feature_Engineering_for_Classification

[5] W. Fan, E. Zhong, J. Peng, O. Verscheure, K. Zhang, J. Ren, R. Yan and Q. Yang, "Generalized and heuristic-free feature construction for improved accuracy," SIAM International Conference on Data Mining, 2010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3085258/

[6] U. Khurana, D. Turaga, H. Samulowitz and S. Parthasarathy, "Cognito:Automated Feature Engineering for Supervised Learning," IEEE 16th International Conference on Data Mining Workshops, 2016. https://ieeexplore.ieee.org/document/7836821

[7] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors, DSAA, 2015. http://www.jmaxkanter.com/static/papers/DSAA_DSM_2015.pdf

[8] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE TPAMI, Vol. 35 No. 8, 2013. http://www.iro.umontreal.ca/~lisa/pointeurs/TPAMISI-2012-04-0260-1.pdf

[9] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artificial Intelligence 97, pp. 245–271, 1997. https://core.ac.uk/download/pdf/82009906.pdf

[10] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research 3, pp. 1157-1182, 2003. http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

[11] O. Janković, "Data Mining: Implikacije filterovanja rangiranjem na performanse klasifikacionog modela," ETRAN 2016, Zlatibor, pp. RT5.5.1-6, 2016.

[12] O Janković, "Data Mining: Implikacije wrapper pristupa selekcije atributa na performanse klasifikacionog modela," XV Međunarodni naučno-stručni simpozijum INFOTEH Jahorina 2016, vol 15, pp.660-664, Mart 2016

[13] J. A. P. Pardo, "Automatic learning procedures for non-invasive blood pressure measurements, " Master thesis, 2009. http://bibing.us.es/proyectos/abreproy/11818/fichero/Portada.pdf

ABSTRACT

The need to have available datasets with features that will allow better predictive performance from one, and less complex and efficient models on the other hand, is an essential part of the compromise and the constant challenge of the predictive modeling process. The paper will show the selection of characteristics, using the unsupervised learning technique - the analysis of the main components (the output options - a subset of characteristics that offer most information), and the impact of dimensional reduction on the performance of prediction of the aircraft engine failure for data driven maintenance model.

Feature engineering:Implications of the feature selection on the performance of prediction

Olivera Janković, Đorđe Babić