Inženjering karakteristika u kontekstu predikcije korištenjem regresije

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

Sažetak— **U fokusu ovoga rada je proces inženjeringa karakteristika sa ciljem obezbjeđenja trening podataka koji posjeduju atribute koji imaju veću prediktivnu snagu za ciljnu predikciju, koji potencijalno omogućavaju bolje performanse prediktivnih modela. U okviru eksperimentalne postavke biće prikazan jedan od načina, sa namjerom da se postigne više sa podacima kojima se raspolaže, za potrebe kreiranja prediktivnih modela - predikcije korištenjem modela regresije, primjenom algoritma linearne regresije, algoritma k-najbliži susjed i višeslojni perceptron klasifikatora, nad podacima koji predstavljaju simulaciju degradacije avionskog motora.**

Ključne riječi - mašinsko učenje; inženjering karakteristika; predikcija; regresija;

I. UVOD

Prediktivno modeliranje je potencijalno veoma koristan te stoga veoma cijenjen dio područja mašinskog učenja. Prediktivni modeli se obučavaju na tzv. istorijskim podacima a donose predikciju na novim podacima. Kako je cilj postići što bolje performanse to je i glavno pitanje kako se mogu postići što bolji rezultati. Jedan od aspekata (često najkorisniji) koji se smatra važnim za podizanja performansi problema mašinskog učenja je poboljšanje performansi sa fokusom na podatke. Naime, zavidni rezultati se mogu očekivati sa promjenama napravljenim na trening podacima i samom definisanju problema. Strategija se u biti ogleda kroz kreiranje novih i različitih perspektiva podataka koji su nam na raspolaganju sa ciljem da se što bolje razotkrije struktura osnovnog problema prema algoritmima mašinskog učenja [1].

Postoje različiti načini poboljšanja performansi prediktivnog modela [2] sa osvrtom na podatke kao što su pored ostalih i pribavljanja većeg obima i boljeg kvaliteta podataka, proces čišćenja podatka, promjena veličine distribucije, selekcija atributa, te kreiranje novih karakteristika – taktika koja je u fokusu ovoga rada. U osnovi je cilj kreirati i dodati nove karakteristike, bilo dekompozicijom atributa na više novih vrijednosti ili agregacijom atributa na primjer, procesom poznatim kao inženjering karakteristika (*feature enginering*).

Cio postupak se ogleda kroz krajnji rezultat procesa inženjeringa karakteristika, kroz mnoštvo novih gledišta i verzija raspoloživog seta podataka (podaci se u tom kontekstu mogu smatrati kao svojevrsne informacije o posmatranom problemu), a vrijednost svakog od odabranih nadalje je neophodno utvrditi i procjeniti primjenom prediktivnih algoritama modeliranja (sam postupak je iterativan, podrazumjeva eksperimentisanje i traži vrijeme). U okviru ovoga rada, korištenjem odgovarajuće eksperimentalne postavke, za ilustraciju odabranog postupka inženjeringa karakteristika i uticaja kreiranih (derivisanih) karakteristika na rezultate predikcije korištenjem regresionog modela (primjenom algoritma linearne regresije, algoritma k-najbližih susjeda i višeslojni perceptron - klasifikatora Weka alata), će se koristiti javno dostupan set podataka koji predstavlja simulaciju degradacije avionskog motora.

II. MAŠINSKO UČENJE I INŽENJERING KARAKTERISTIKA

Mašinsko učenje [3], kao svojevrstan vid potrage za obrascima u podacima koji nas okružuju (putem algoritama i računarskih sistema), jedan je od načina koji može da se koristi za kreiranje prognostičkih/prediktivnih modela. Značajan faktor uspjeha projekata mašinskog učenja svakako su korištene karakteristike. Dosta često, sirovi podaci nisu u obliku koji je pogodan za učenje, ali se iz njih mogu konstruisati karakteristike koje to jesu. Inženjering karakteristika je proces pretvaranja sirovih podataka u karakteristike koje bolje predstavljaju temeljni problem prediktivnih modela, sa ciljem veće tačnosti modela na neviđenim podacima.

Inženjering karakteristika [4] je tema koja nije dovoljno obrađivana iako se cijeni da je od vitalnog značaja za uspjeh mašinskog učenja, odnosno da je veliki dio uspjeha mašinskog učenja zapravo uspješnost u inženjeringu karakteristika. Potrebne su takve karakteristike koje opisuju strukture svojstvene korištenim podacima.

Algoritmi mašinskog učenja nauče rješenje za problem korištenjem podataka uzorka za trening. Međutim, podaci su pri svakom novom projektu različiti i u biti praksa i empirijsko obučavanje se neophodni u postizanju dobrih odabira pri odlučivanju koje postupke i kada odabrati. Ovladavanje procesom inženjeringa karakteristika dolazi sa vlastitom praksom i iskustvom i proučavanjem onoga što su radili drugi koji su u tome postigli dobre rezultate.

Karakteristike je u biti potrebno ručno kreirati (moderne metode dubinskog učenja (*deep learning*) postižu određeni

uspjeh u području automatske identifikacije i korištenje karakteristika na sirovim podacima; npr. korištenjem restriktivne Bolcmanove mašine - tip vještačkih neuronskih mreža) što posledično zahtjeva mnogo vremena za upoznavanje i razumjevanje stvarnih podataka uzorka (ne njihovih agregata). Potrebno je poimanje o osnovnom obliku problema, strukturi u podacima i kako najbolje da ih izlože prediktivnim algoritama modeliranja.

Primjeri manuelnih konstrukcija karkteristika koji se mogu primjeniti u praksi su i dekompzicija kategoričkih atributa (poprimaju određen broj diskretnih vrijednosti) i dekompozicija datum_vrijeme forme. U kontekstu pak tabelarnih podataka, to često znači mješavinu agregiranja ili kombinovanja karakteristika za kreiranje novih karakteristika, i dekomponovanja ili dijeljenje karakteristika da bi se kreirale nove karakteristike. U situacijama kada se od istorijskih podataka koji dolaze sa vremenskom oznakom (*time stamp*), inicirajući tako vrijeme prikupljanja za svaki dio podataka pojednačnom, pripremaju setovi podataka za trening veoma je česta potreba da se kreiraju i nove karakteristike. Postoje razni načini kreiranja karakteristika od podataka koji dolaze sa podacima sa vremenskim oznakama, pri čemu mjerne jedinice za vrijeme mogu biti sekunde, minute, sati, dani, milje, ciklusi,... U kontekstu takvog problema veoma su važne tzv. lag karakteristike (lag, usporavanje, kašnjenje, pomaknuto ..), koje treba konstruisati iz izvora podataka koji dolaze sa vremenskom oznakom.

U okviru ovog rada biće korištena jedna od opštih tehnika kotrljajućih agregata (*rolling aggregate*) u kojoj se za svaki zapis opreme bira veličina kotrljajućeg (rolling) prozora veličine "W" koji predstavlja broj jedinica vremena za koji se želi izračunati istorijski agregati [5]. Nakon toga se izračunavaju karakteristike kotrljajućih agregata koristeći W periode prije datuma /vremenske oznake tog zapisa. U okviru date eksperimentalne postavke biće korišteni primjeri kotrljajućih agregata srednja vrijednost i standardna devijacija (pojedinačno i zajedno).

III. Eksperimentalne postavke i rezultati

A. *Ulazni skup podataka*

Ulazni skup podataka "Turbofan Engine Degradation Simulation Data Set", korišten u okviru eksperimentalne postavke, predstavlja simulaciju degradacije avionskog motora (upotrebom C-MAPSS, aero-pogon simulator sistema) doniran od strane NASA [6]. U biti su simulirana četiri različita seta, u različitim kombinacijama pogonskih uslova i modova kvara, pri čemu se snima određen broj senzorskih kanala za karakterizaciju evolucije greške.

Konkretnije, u okviru postavke je korišten jedan od četiri simulirana seta podataka iz originalnog seta podataka, pri čemu se svaki od njih sastoji od tri seta podataka – podataka za trening, testiranje i referentih podataka. Za obučavanje u okviru ove ekperimentalne postavke iz pomenutog originalnog seta podataka biće korišten set podataka "Train_FD003.txt", koji sadrži trening podatke koji predstavljaju podatke avionskog motora radom do otkaza (*run-to–failure*) [7]. Podaci su dati

```
1)    unit number
2)    time, in cycles
3)    operational setting 1
4)    operational setting 2
5)    operational setting 3
6)    sensor measurement  1
7)    sensor measurement  2
...
26)   sensor measurement  26
```

Slika 1. Prikaz kolona originalnog seta podataka

kao zip-kompresovana tekstualna datoteka sa 26 kolona brojeva, razdvojenih razmacima pri čemu svaki red predstavlja snimak (*snapshot*) podataka uzetih tokom jednog operativnog ciklusa, svaka kolona je drugačija promenljiva (izvorni prikaz kolona dat je na Sl. 1).

Set podataka se u osnovi sastoji od višestrukih multivarijantnih vremenskih serija. Svaki put podaci vremenskih serija su serija je iz drugog motora pri čemu se smatra da pripadaju floti motora iste vrste. Podrazumjeva se da svaki motor počinje sa različitim stepenom početnog habanja i proizvodnih varijacija koja je nepoznata za korisnika. Ovo habanje i varijacije se smatraju normalnim, odnosno ne smatraju se stanjem neispravnosti. Postoje tri operativne postavke koje imaju značajan uticaj na performanse motora koje su uključene u podacima. U ovim simuliranim podacima, podrazumjeva se da na početku svake serije motor radi normalno i da degradacija rada počinje da se dešava u nekoj tački serije operativnog ciklusa i njen progres raste u intenzitetu. Kada se dosegne predefinisani prag (*threshold*) tada se motor smatra nepouzdanim za buduće operacije. U kontekstu ovog seta podataka zadnji ciklus se može smatrati tačkom kvara za određeni motor (npr. motor sa oznakom 1 je pretrpio neuspjeh u 259-om cikusu).

```
@relation Serije FD003
@attribute ID_motora numeric
@attribute Ciklus numeric
@attribute Setovanje1 numeric
@attribute Setovanje2 numeric
@attribute Setovanje3 numeric
@attribute Senzor1 numeric
@attribute Senzor2 numeric
@attribute Senzor3 numeric
@attribute Senzor4 numeric
@attribute Senzor5 numeric
@attribute Senzor6 numeric
@attribute Senzor7 numeric
@attribute Senzor8 numeric
@attribute Senzor9 numeric
@attribute Senzor10 numeric
@attribute Senzor11 numeric
@attribute Senzor12 numeric
@attribute Senzor13 numeric
@attribute Senzor14 numeric
@attribute Senzor15 numeric
@attribute Senzor16 numeric
@attribute Senzor17 numeric
@attribute Senzor18 numeric
@attribute Senzor19 numeric
@attribute Senzor20 numeric
@attribute Senzor21 numeric
@attribute Preostali ciklus numeric

@data
1,1,-0.0005,0.0004,100,518.67,642.36,1583.23,1396.84,14.62,21.61,553.97,
2387.96,9062.17,1.3,47.3,522.31,2388.01,8145.32,8.4246,0.03,391,2388,100
,39.11,23.3537,258
```

Slika 2. Izgled originalnog, prilagođenog .arff fajla (bez agregiranih karakteristika) nakon označavanja, za trening regresionih modela predikcije (Označeni primjer je primjer iz skupa podataka za kojeg je poznata vrijednost ciljne varijable/atributa)

Za testiranje su korišteni testni podaci Test_FD003.txt, koji imaju istu šemu podataka kao i trening podaci pri čemu razlika leži u tome da oni ne sadrže podataka kada se desio kvar. Set podataka RUL_FD003.txt sadrži referentne podatke (u biti jednodimenzionalna tabela, vektor), koji obezbjeđuju broj preostalih radnih ciklusa za motor koji se nalazi u testnim podacima.

Karakteristike koje će biti uključene u trening podacima mogu biti grupisane u dvije kategorije. Selektovani su svi sirovi atributi, atributi koji su uključeni u originalne ulazne podatke. Pored njih tu su i agregirani atributi, atributi koji u osnovi sažimaju istorijske aktivnosti za posmatranu problematiku vezanu za avionski motor. Konkretno, u okviru eksperimentalne postavke ovoga rada kreirana su setovi dva tipa agregiranih karakteristika za svaki od 21 senzora podataka sa Sl.2:

- SSenzor1–SSenzor21: predstavljaju kretanje prosječne vrijednosti – srednja vrijednost sa senzora u najviše W prethodnih ciklusa
- SDSenzor1-SDSenzor21: standardna devijacija senzorskih vrijednosti u najviše W nedavnih ciklusa

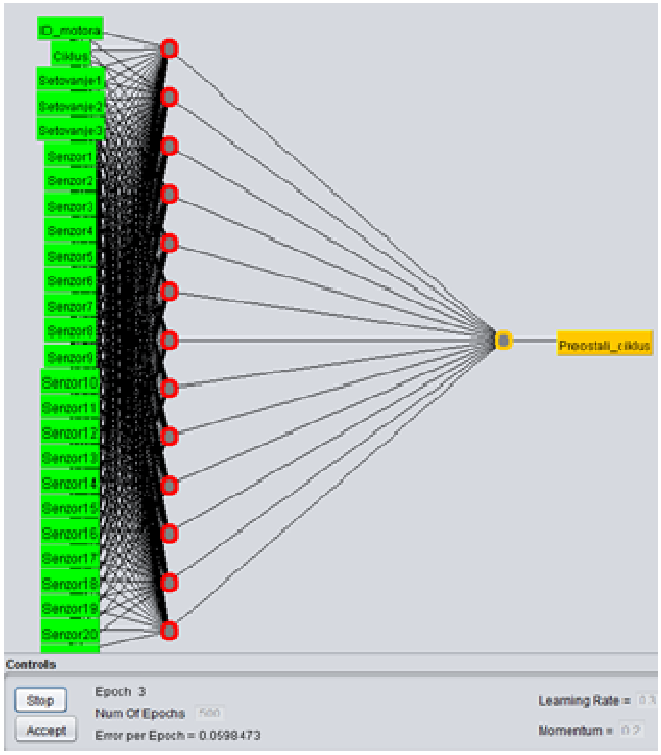
a koji su uključeni u trening skup podataka pojedinačno i zajedno. Naime, kreirana su tri trening skupa podataka, koji predstavljaju kombinacije: sirovih podataka i SSenzor1–SSenzor21 (označeno sa SS), sirovih podataka i SDSenzor1-SDSenzor21 (označeno sa SDS), sirovih podataka i SSenzor1–SSenzor21 i SDSenzor1-SDSenzor21(označeno sa SSiSDS). Posmatrano sumarno kroz broj atributa to znači 21, 21 i 42 dodatna atributa u odnosu na početni (originalni) koji ima 26 atributa, odnosno 48, 48, 69 atributa ukupno za pomenuta tri trening skupa respektivno (u zbir uračunat i atribut klase koji je naknadno kreiraran u skladu sa ciljem predikcije). Analogan postupak se odvija i na testnim podacima. Eksperimentalna postavka uključuje i dvije vrijednosti parametra vremenskog prozora (W=5 i W=10), za koje se vrši agregiranje, za sva tri navedena trening skupa.

Predikcija putem regresije korištena u ovom radu, u kontekstu prirode izabranog seta podataka, se odnosi na predviđanje kvara avionskog motora, odnosno predviđanje preostalog broja ciklusa koje će motor imati prije nego se desi kvar (ciklus je u tom smislu mjerna jedinica vremena). Neophodan korak (već je pomenuto da trening skup (niti testni analogno tome) izvorno ne sadrže oznaku instanci) da bi se izvršila predikcija putem regresije je označavanje uzoraka. Oznaka uzoraka (*label*) za potrebe regresije, (prethodno kreiran atribut Preostali ciklus (Sl.2)), poprima vrijednosti koje predstavljaju preostali broj ciklusa za svaki od primjera pojedinačno. Sa Sl.2. se može vidjeti da vrijednost za Preostali ciklus, za ID_motora=1 i Ciklus=1, iznosi 258 (obzirom da ukupan broj ciklusa za ID_motora=1 iznosi 259; po analogiji za ID_motora=1 i Ciklus=259, vrijednost Preostali ciklus=0). Označavanje testnih podataka Test_FD003.txt je izvršeno na bazi referentnih podataka RUL_FD003.txt, pri čemu, analogno skupu trening podataka, prolaze kroz istu proceduru označavanja primjera.

```
Preostali_ciklus =

    -0.2377 * ID_motora +
    -0.2818 * Ciklus +
    -7.8446 * Senzor2 +
    -0.739  * Senzor3 +
    -0.6959 * Senzor4 +
  -525.5756 * Senzor6 +
   -78.2624 * Senzor8 +
    -0.315  * Senzor9 +
   473.1422 * Senzor10 +
   -38.2813 * Senzor11 +
     1.0287 * Senzor12 +
  -106.9698 * Senzor13 +
    -0.323  * Senzor14 +
   -40.5947 * Senzor15 +
    -2.981  * Senzor17 +
    18.9456 * Senzor20 +
    30.8122 * Senzor21 +
 467276.2501
```

Slika 3. Predikcioni model linearne regresije za originalni set trening podataka



Slika 4. Neuronska mreža sa 13 neurona (srednja, cjelobrojna vrijednost sume broja ulaznih i izlaznih čvorova (26+1)/2)), za originalni skup podataka

Za potrebe izgradnje regresionog modela korištena su tri različita klasifikatora: algoritam linearne regresije (Linear Regresion), algoritam k najbližeg susjeda (IBk) i višeslojni

perceptron (MLP, Multilayer Perceptron). Linearna regresija je vjerovatno jedan od najpoznatijih i shvaćenih algoritama u mašinskom učenju. Obučavanje modela linearne regresije znači procjenu vrijednosti koeficijenata koji se koriste u prikazu s podacima koje imamo na raspolaganju (ilustracija jednog od modela na Sl.3).

U eksperimentu je korišten i IBk klasifikator koji predstavljaja algoritam k najbližeg susjeda. Prosti algoritam najbližeg susjeda IB1 koristi normalizovano Euklidsko rastojanje (najkraća udaljenost između dvije tačke u jednom prostoru), dok algoritam k najbližeg susjeda IBk može selektovati odgovarajuću vrijednost za k baziranu na kros validaciji. Višeslojni perceptroni predstavljaju najistaknutiji tip vještačke neuronske mreže i pripadaju klasi mreža sa prostiranjem unaprijed, tzv. feedforvard mreže, koje ne sadrže nikakve cikluse (opozit su periodične neuronske mreže koje imaju cikluse). U radu će biti korištena implementacija MLP Weka alata [8]. Ova funkcija implementira algoritam povratnog prostiranja (backpropagation) za izgradnju modela neuronske mreže (Sl. 4) za klasifikaciju instance. U Tabeli I. II i III su prikazani postignuti rezultati za svaki korišteni klasifikator pojedinačno.

Za potrebe kreiranje pomenutih verzija trening i testnih setova podataka (ukupno po sedam, originalna i različite opcije agregiranih karakteristika i vremenskog okvira) i procesa označavanja klase u skladu sa odabranom vrijednošću predviđanja razvijena je adekvatna programska podrška (C#), a tako kreirani trening i testni skupovi korišteni su u okviru Weka alata (.arff formalizovan Weka format) za data mining - korištenjem izvedbe Weka alata prethodno pomenutih klasifikatora [9].

Procjena performansi izgrađenog modela je krajnji, potreban postupak. Najjednostavnije, procjena bi u biti mogla biti performanse modela korištenjem trening podataka, procjena koja je zasigurno optimistična, obzirom da je model krojen/obučavan na podacima koji se ujedno i koriste za procjenu performanse. U suštini, neophodan je robustniji model, više realan pristup za mjerenje performansi modela (procjena na novim, prethodno nekorištenim podacima). Pored toga, u slučajevima, opcijama koje sadrže vremenski označene podatke [10][11], tipične trening i testne rutine trebaju da uzmu u obzir aspekte vremenskih promjena (npr. tako da su svi test primjerci kasnije u vremenu, u odnosu na trening i validacione primjerke) stoga uobičajeni postupci korištenjem k-struke unakrsne validacija se ne preporučuju. Procjena performansi kreiranih prediktivnih modela u ovom radu je izvršena opcijom testiranja korištenjem prethodno pomenutih testnih podataka (prethodno pripremljenih u skladu sa pripremama izvršenim na trening podacima (neophodnost iste šeme; jedan originalni i dodatnih šest različitih skupova testnih podataka)).

U Tabeli I,II i III su prikazani postignuti rezultati različitih klasifikatora linearne regresije, k-najbližih susjeda IBk i MLP respektivno, dobijeni treniranjem nad kompletnim setovima za trening i evaluacijom korištenjem 100 testnih zapisa - za svaki motor po jedan maximalni ciklus (postoji opcija sa kompletnim setom podataka, kada se koriste sve dostupne vremenske serije

TABELA I. EKSPERIMENTALNI REZULTATI REGRESIJE KORIŠTENJEM ALGORITMA LINEARNE REGRESIJE ZA ORIGINALNI SET I SETOVE PODATAKA NASTALE KREIRANJEM AGREGIRANIH KARAKTERISTIKA ZA DVIJE VRIJEDNOSTI VREMENSKOG OKVIRA W (EVALUACIJA KORIŠTENJEM 100 TESTNIH INSTANCI)

	Originalni set podataka	**W=5**			**W=10**		
		SS	***SDS***	***SS i SDS***	***SS***	***SDS***	***SS i SDS***
Koeficijent korelacije	0.7991	0.7968	0.7955	0.792	**0.8009**	**0.808**	**0.811**
MAE	46.2731	**43.1519**	**44.9926**	**43.0607**	**42.26**	**42.4305**	**39.625**
RMSE	57.3835	**52.8533**	**56.1876**	**53.1716**	**52.1778**	**52.8885**	**49.8136**
RAE (%)	73.1348	**68.2018**	**71.1111**	**68.0576**	**66.7922**	**67.0616**	**62.6275**
RRSE (%)	76.3281	**70.3023**	**74.7374**	**70.7256**	**69.4037**	**70.3491**	**66.259**

TABELA II. EKSPERIMENTALNI REZULTATI REGRESIJE KORIŠTENJEM ALGORITMA IBK (K=20), ZA ORIGINALNI SET I SETOVE PODATAKA NASTALE KREIRANJEM AGREGIRANIH KARAKTERISTIKA ZA DVIJE VRIJEDNOSTI VREMENSKOG OKVIRA W (EVALUACIJA KORIŠTENJEM 100 TESTNIH INSTANCI)

	Originalni set podataka	**W=5**			**W=10**		
		SS	***SDS***	***SS i SDS***	***SS***	***SDS***	***SS i SDS***
Koeficijent korelacije	0.7829	**0.7944**	**0.8083**	**0.8075**	**0.8121**	**0.8018**	**0.7999**
MAE	31.2767	**29.7719**	35.4943	34.1414	**29.0933**	34.6262	33.1795
RMSE	47.1995	**45.1384**	48.3893	48.5635	**42.4314**	49.2504	49.144
RAE(%)	49.4329	**47.0547**	56.0989	53.9607	**45.9822**	54.7269	52.4404
RRSE(%)	62.7819	**60.0404**	64.3645	64.5963	**56.4397**	65.5098	65.3683

TABELA III. EKSPERIMENTALNI REZULTATI REGRESIJE KORIŠTENJEM ALGORITMA MULTILEJER PERCEPTRON, ZA ORIGINALNI SET I SETOVE PODATAKA NASTALE KREIRANJEM AGREGIRANIH KARAKTERISTIKA ZA DVIJE VRIJEDNOSTI VREMENSKOG OKVIRA W (EVALUACIJA KORIŠTENJEM 100 TESTNIH INSTANCI)

	Originalni set podataka	**W=5**			**W=10**		
		SS	***SDS***	***SS i SDS***	***SS***	***SDS***	***SS i SDS***
Koeficijent korelacije	0.756	**0.793**	**0.8547**	**0.8232**	0.751	**0.7945**	**0.7899**
MAE	30.7872	**28.115**	33.9339	34.0104	31.6609	**30.4008**	31.269
RMSE	42.563	**38.4267**	**38.9962**	44.713	**41.4782**	**41.1081**	**38.0084**
RAE(%)	48.6593	**44.4359**	53.6327	53.7537	50.0402	**48.0486**	49.4208
RRSE(%)	56.6147	**51.1129**	**51.8704**	59.4745	**55.1718**	**54.6795**	**50.5565**

testnog seta). Svi rezultati koji su bolji od onih postignutih nad originalnim setom podataka, za svaki klasifikator pojedinačno, su boldirani, a najbolje vrijednosti za svaku korištenu mjeru su označene crvenom bojom.

Fitnes dobijenih modela, kao što se može vidjeti u pomenutim tabelama dati su sledećim (Weka dostupni) statističkim parametrima: koeficijent korelacije (*Correlation coefficient*) kao specifičan pokazatelj reprezentativnosti regresije (vrijednosti u rasponu od 0 do 1, reprezentativnija veća vrijednost), srednja apsolutna greška MAE (*mean absolute error*), korijen srednje kvadratne greške RMSE (*root mean square error*), relativna apsolutna greška RAE (*relative absolute error*), korijen relativne kvadratne greške RRSE (*root relative squared error*); pri čemu se najčešće za usporedbu performansi prediktivnih modela koriste vrijednosti srednje apsolutne greške MAE i korijena srednje kvadratne greške RMSE.

Na osnovu obavljenih eksperimenata, dobijenih i prikazanih rezultata neki od baznih zaključaka koji se mogu izvesti su sledeći:

- Svaki od korištenih klasifikatora imao je određena poboljšanja, koristi od najmanje jedne od šest kreiranih opcija setova podataka (sa kreiranim karakteristikama, za određene lag vrijednosti).
- Različiti su skupovi trening podataka sa kojima su klasifikatori postigli najbolje rezultate posmatrano za svaki klasifikator pojedinačno u odnosu na pripadajuće originalne trening skupove podataka. Npr. klasifikator linearne regresije postigao je najbolji rezultat na trening skupu podataku u kojem se nalaze SS i SSD karakteristike zajedno, dok se najbolji rezultat za IBk klasifikator nad trening skupom podataka u kome se nalaze SS karakteristike, oba po svim mjerama korištene metrike (vrijednosti su označene crvenom bojom).
- Vidljivo je da je na rezultate klasifikatora uticala i odabrana veličina W za isti tip načelno odabranih agregiranih karakteristika, pri čemu su u u dva, prethodno pomenuta klasifikatora (linearna regresija i IBk) postignuti bolji rezultati za opciju većeg vremenskog okvira (W=10).
- Klasifikator višeslojni perceptron je postigao najbolji rezultat pojedinačno za svaku korištenu mjeru ali pojedinačno posmatrano (uzimajući u obzir sve trening skupove i korištene klasifikatore), npr. najmanja vrijednost srednje apsolutne greške 28.115, najmanja vrijednost korijena srednje kvadratne greške 38.0084.

IV. ZAKLJUČAK

U okviru rada prikazan je jedan od načina kreiranja karakteristika procesa inženjeringa karakteristika, sa ciljem da se, u okviru eksperimentalnih postavki, ilustruje navedena opcija i da se pokaže, obzirom da apriori odgovora nema, u kojoj mjeri navedene karakteristike doprinose boljim performansama prediktivnih modela primjenom regresije (korištenjem tri različita klasifikatora). Postignuti rezultati ukazuju na pozitivan uticaj primjenjene opcije na primjeru korištenih klasifikatora u mjeri koja potvrđuju da pravila nema te da je u biti neophodno probati različite nivoe agregacije i izvršiti evaluaciju performansi modela u cilju određivanja optimalnog nivoa agregacije. Smjernice za daljnja istraživanja u kontekstu prikazanog mogu se odnositi na primjenu drugih opcija agregiranja karakteristika za iste ili različite vrijednosti vremenskog okvira na primjer.

LITERATURA

[1] J. Heaton, "An empirical analysis of feature engineering for predictive modeling", SoutheastCon 2016, pp. 1- 6, 2016.

[2] R. Boire, Feature Engineering within the Predictive Analytics Process — Part One, http://www.predictiveanalyticsworld.com/patimes/feature-engineering-within-the-predictive-analytics-process-part-one/7657/2016.

[3] P. Domingos, "A Few Useful Things to Know about Machine Learning", Magazine Communications of the ACM, Volume 55 Issue 10, pp 78-87, October 2012.

[4] O. Janković, "Inženjering karakteristika u kontekstu predikcije korištenjem binarne klasifikacije", YUINFO 2017, prihvaćen za objavljivanje

[5] F. B. Uz, Predictive Maintenance Modelling Guide R Notebook, Cortana Intelligence Gallery, 2016, https://gallery.cortanaintelligence.com/Notebook/Predictive-Maintenance-Modelling-Guide-R-Notebook-1 2016.

[6] A. Saxena and K. Goebel "Turbofan Engine Degradation Simulation Data Set", NASA Ames Prognostics Data Repository, 2008 (http://ti.arc.nasa.gov/project/prognostic-data-repository), NASA Ames Research Center, Moffett Field, CA, 2008.

[7] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "DamageRun-to-Failure Simulation", in the Proceedings of the Ist International Conference on Prognostics and Health Management (PHM08), Denver CO, Oct 2008.

[8] O. Janković, "Primjena i evaluacija klasifikatora višeslojnog perceptrona za potrebe klasifikacije linearno neseparabilnih problema", SYM-OP-IS 2015, str. 208-211, Ivanjica, 2015.

[9] I. H. Witten, E. Frank. Data Mining Practical Machine Learning Tools and Techniques, Elsevier, 2011.

[10] O. Janković, "Data Mining: Evaluacija klasifikacije iz perspektive strima podataka", YUINFO 2016, str. 373-378, Kopaonik, 2016.

[11] O. Janković, "Modeliranje i predikcija podataka vremenskih serija u kontekstu data mininga", VII Naučni skup MREŽA 2015, Valjevo, Zbornik Radova, str. 41-47

ABSTRACT

The focus of this work is a process engineering characteristics with the aim of ensuring the training data that have the attributes that have greater predictive power for prediction of the target, which enable better performance of predictive models. Within the experimental setting will be shown one way, with the intention to achieve more with available data, for the purposes of creating predictive models - prediction model using regression algorithm using the linear regression algorithm, k-nearest neighbor and multilayer perceptron classifier, using data that represent a simulation of the aircraft engine degradation.

ENGINEERING CHARACTERISTICS IN THE CONTEXT OF PREDICTION USING REGRESSION

Olivera Janković