

Analiza podataka o mašinskom učenju i istraživanju podataka

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Petar Mičić
micicpetar73@gmail.com

16. avgust 2019

Sažetak

U svetu postoji sve veće interesovanje za istraživanje, obradu, rukovanje podacima u različite svrhe, stoga je ovo primer rada u kome je izvršeno istraživanje podataka svih onih koji imaju iskustva iz oblasti mašinskog učenja i istraživanja podataka. Preuzeti su prikupljeni podaci o programskim jezicima, alatima, bibliotekama, okruženjima, skladištima podataka, alatima za prikaz podataka, proizvodima mašinskog učenja, i drugi. Adekvatnim pretprocesiranjem, primenom najpoznatijih algoritama iz oblasti istraživanja podataka prikazani su različiti zanimljivi rezultati, kao rezultat njihove primene.

Sadržaj

1	Uvod	2
2	Podaci	2
2.1	Statistike	2
2.2	Nedostajuće vrednosti	3
3	Klasterovanje	4
3.1	Preprocesiranje podataka	4
3.2	Klasterovanje koriscenjem KModes algoritma	4
3.2.1	Klasterovanje prema okruženjima	4
3.2.2	Klasterovanje prema bibliotekama	5
3.3	Klasterovanje koriscenjem algoritma K-Sredina	5
3.3.1	Klasterovanje prema visini obrazovanja i industriji	5
3.3.2	Klasterovanje prema zaradama	6
3.4	Klasterovanje Kohenen algoritmom	6
3.4.1	Klasterovanje prema jezicima i visini obrazovanja	6
4	Zaključak	7
5	Literatura	7
	Literatura	7

1 Uvod

Oktobra 2018. godine je postavljena anketa o mašinskom učenju i istraživanju podataka koja je bila aktuelna jednu nedelju. Svrha ankete je da prikaže pravi pogled na svet mašinskog učenja i istraživanja podataka. Nakon prečišćavanja podataka, sakupljeno je 23 859 odgovora. U ovom radu će uz pomoć programskog jezika Python i SPSS Modeler alata biti istraživani prikupljeni podaci. Bice izvršeno klasterovanje korišćenjem algoritama K-Means, K-Modes, Kohonen i ROCK. Cilj ovog rada jeste da prikaže kako se vrši klasterovanje različitim algoritmima i da budu prikazane razlike dobijenih rezultata.

2 Podaci

Ispitanici su dobili zadatak da odgovore na 50 pitanja. Veliki broj pitanja su selektivnog tipa, gde se od ispitanika očekivalo da obeleže više ponudjenih odgovora (eng. mutiple choice responses). Kao takvi, podaci se uglavnom sastoje od kategoričkih i binarnih atributa. Primeri pitanja iz ankete:

- odaberite programske jezike koje koristite redovno
- odaberite programske jezike koje biste preporučili da nauče istraživači podataka
- odaberite biblioteke mašinskog učenja koje ste koristili u proteklih 5 godina
- odaberite biblioteke ili alate za vizuelni prikaz podataka koje ste koristili u proteklih 5 godina
- odaberite skladišta podataka koja ste koristili prilikom istraživanja podataka u proteklih 5 godina

Odgovori ispitanika su smešteni u datoteku mutipleChoiceResponses.csv. Datoteka sadrži 23859 redova i 396 kolona. Podaci se mogu preuzeti na sledecem [linku](#). Tabela 1 prikazuje deo prikupljenih podataka.

Q2	Q4	Q6	Q7	Q8
45-49	Doctoral degree	Consultant	Other	
30-34	Bachelor's degree	Other	Manufacturing/Fabrication	5-10
30-34	Master's degree	Data Scientist	I am a student	0-1
35-39	Master's degree	Not employed		
22-24	Master's degree	Data Analyst	I am a student	0-1
25-29	Bachelor's degree	Data Scientist	Computers/Technology	0-1
35-39	Doctoral degree	Other	Academics/Education	10-15
18-21	Master's degree	Other	Other	0-1
25-29	Master's degree	Not employed		
30-34	Master's degree	Software Engineer	Internet-based Services	3-4

Tabela 1: Deo prikupljenih podataka

2.1 Statistike

Sledi prikaz nekoliko statistickih podataka o ispitanicima:

- Ukupan broj ispitanika iznosi 23859, medju kojima je 19430 muških (81.44%) i 4010 ženskih ispitanika (16.81%), pri čemu se suzdržalo da se izjasni 340 ispitanika (1.34%), a 79 (0.33%) bi se opisalo samostalno.
- Najveci broj ispitanika je starosti izmedju 25 i 29 godina iznosi 3646 (28.07%), broj ispitanika starosti od 22 do 24 godine iznosi 2575 (19.82%), a broj ispitanika starosti od 30 do 34 godine iznosi 2219 (17.08%).
- Jezik koji se najčešće koristi u mašinskom učenju je Python koji je označilo 52,07% ispitanika, zatim R koji je označilo 36,90% ispitanika.
- Najpopularniji jezik u mašinskom učenju je Python koji koristi 11959 ispitanika, a potom SQL koji koristi 5832 ispitanika, dok R koristi 4806 ispitanika.
- Broj ispitanika sa zaradom od 0 do 10,000 dolara na godišnjem nivou iznosi 2,293, dok godišnju zaradu izmedju 400,000 i 500,000 dolara ima 11 ispitanika.
- Medju ispitanicima najviše je onih sa godinu dana iskustva iz oblasti mašinskog učenja (3,283), a 99 sa preko 30 godina iskustva.

2.2 Nedostajuće vrednosti

U podacima je bilo nedostajućih vrednosti. Atribut koji se odnosi na najkorišćenije biblioteke za prikaz podataka ima najviše nedostajućih vrednosti, čak 11675 (48.931%). Kompletan prikaz nedostajućih vrednosti se može videti na [1](#).

Field	Measurement	% Complete	Valid Records	White Space
edu level	Categorical	98.236	23439	421
undergraduate major	Categorical	96.178	22948	912
current role	Categorical	95.981	22901	959
current industry	Categorical	90.889	21686	2174
xp	Categorical	88.441	21102	2758
salary	Categorical	84.602	20186	3674
ml methods in business	Categorical	86.63	20670	3190
most often lang	Categorical	63.801	15223	8637
recommended lang	Categorical	78.747	18789	5071
most often ml lib	Categorical	54.443	12990	10870
data visualization lib	Categorical	51.069	12185	11675
perc of time coding	Categorical	77.737	18548	5312
years of analyzing data	Categorical	77.678	18534	5326
years of ml methods	Categorical	77.502	18492	5368
data scientist	Categorical	77.456	18481	5379
type of data	Categorical	58.168	13879	9981
demonstrates expertise	Categorical	66.555	15880	7980
percent of data projects	Categorical	54.987	13120	10740
percent of projects exploring model	Categorical	55.7	13290	10570
ml models to be black boxes	Categorical	56.031	13369	10491

Slika 1: Udeo ispravnih vrednosti u podacima

3 Klasterovanje

Mnoge aplikacije zahtevaju podelu objekata na intuitivno slične grupe. Podela velikog broja podataka u manje grupe omogućava lakše sumiranje i razumevanje podataka za različit broj aplikacija u istraživanju podataka. Neformalna i intuitivna definicija klasterovanja glasi: *Za dati skup objekata, podeliti objekte u grupe vrlo sličnih objekata*^[1].

S obzirom na to da se među podacima nalazi veliki broj kategoričkih atributa, biće izvršeno klasterovanje algoritmima: K-Modes, K-Means, Kohenen. Klasterovanje će biti usmereno samo na instance koje imaju popunjena polja Python i R (dakle da li ispitanik koristi Python, R ili oba).

3.1 Preprocesiranje podataka

Podaci o mašinskom učenju i istraživanju podataka korišćeni u ovom radu su najvećim delom kategoričkog tipa. Klasterovanje je orjentisano prema različitim vrednostima atributa, što uključuje odsecanje velikog broja podataka. Kako je veliki broj upitnika nepotpun, obrađeni su samo oni ispitanici koji su odgovorili na većinu pitanja.

Za potrebe algoritma K-Sredina, kategoričke vrednosti se implicitno preslikavaju u skup prirodnih brojeva, i nad takvim podacima se primenjuje algoritam.

3.2 Klasterovanje koriscenjem KModes algoritma

3.2.1 Klasterovanje prema okruženjima

Klasterovanje orjentisano jezicima Python i R, pri čemu se ispitanici grupišu prema okruženjima koje koriste tokom istraživanja podataka kao i alata za vizuelizaciju podataka. Klasterovanje je izvršeno algoritmom KModes u programskom jeziku Python. Spisak svih klastera prikazan je u tabeli 2

c0	Python & R	Jupyter/IPython	RStudio	PyCharm	ggplot2	Matplotlib	Plotly	Seaborn
c1	Python & R	Jupyter/IPython	RStudio		ggplot2	Matplotlib	Plotly	Seaborn
c2	Python	Jupyter/IPython				Matplotlib		
c3	R		RStudio		ggplot2			
c4	Python & R	Jupyter/IPython	RStudio		ggplot2	Matplotlib		
c5	Python	Jupyter/IPython		PyCharm		Matplotlib	Plotly	Seaborn

Tabela 2: Reprezentativne vrednosti klastera

Kod klasterovanja ispitanika koji koriste Python, R ili oba programska jezika može se vrlo lako uočiti da oni koji koriste Python uglavnom koriste Jupyter ili IPython, PyCharm kao okruženja, a Matplotlib, Plotly i Seaborn, dok oni koji koriste pored Python-a i R, ili samo R koriste okruženja RStudio i biblioteku ggplot2. S obzirom da su pitanja selektivnog tipa, ukoliko se u opisu klastera ne nalazi vrednost, to znači da većina ne koristi tu biblioteku.

3.2.2 Klasterovanje prema bibliotekama

Grupisanje ispitanika koji regularno koriste Python i R prema alatima za mašinsko učenje. Može se uočiti da većina ispitanika koristi Scikit-Learn biblioteku, dok randomForest koriste oni koji takodje koriste i R programski jezik.

c0	Python	Scikit-Learn	TensorFlow	Keras		Xgboost	
c1	PythonR	Scikit-Learn				Xgboost	randomForest
c2	Python	Scikit-Learn					
c3	Python	Scikit-Learn	TensorFlow	Keras			
c4	PythonR	Scikit-Learn	TensorFlow	Keras	Caret	Xgboost	randomForest
c5	R						

Tabela 3: Reprezentativne vrednosti klastera

3.3 Klasterovanje koriscenjem algoritma K-Sredina

Algoritam K-Sredina kao parametar dobija k, odnosno broj klastera. Svakom klasteru pridružuje se centroid (reprezentativna tačka klastera). Početni centroid se bira na slučajan način. Svaka tačka je dodeljena klasteru sa najbližim centroidom. U svakoj iteraciji centroidi se ažuriraju tako sto dobiju srednju vrednost (eng. mean) svih tačaka u klasteru. Algoritam konvergira do pomenute mere. Vremenska složenost algoritma je $O(n*k*d*i)$, a prostorna $O((n+k)*d)$, pri čemu je n - broj tačaka, k - broj centroida, d - dimenzionalnost a i - broj iteracija.

3.3.1 Klasterovanje prema visini obrazovanja i industriji

Sledi prikaz klasterovanja nad atributima 'Python & R', 'Degree', 'Industry', 'Most often language', 'Current role', koji opisuju redom da li ispitanik regularno koristi Python, R ili oba, stepen obrazovanja ispitanika, industriju u kojoj radi ispitanik, najcesce korisceni jezik i trenutnu njegovu trenutnu poziciju. Klasterovanje je izvršeno algoritmom K-Means, pri čemu senka koeficijent iznosi 0.4. Klasteri su prikazani na slici...

c1	(4127) 32.7%	Python	Python	BSc	Comp/Tech
c2	(3373) 26.7%	Python	Python	MSc	Comp/Tech
c3	(2064) 16.4%	Python	Python	MSc	Comp/Tech
c4	(1334) 10.6%	Python & R	R	MSc	Comp/Tech
c5	(1045) 8.3%	R	R	MSc	Academia
c6	(669) 5.3%	Python & R	Python	Phd	Academia

Tabela 4: Reprezentativne vrednosti klastera

Mogu se uociti 3 klastera. Prvi klaster (c1) cine pretežno studenti koji najcesce koriste programski jezik Python i su završili ili završavaju osnovne studije. Drugi klaster (c5) cine pretežno analiticari koji koriste programski jezik R i rade na akademiji. Treci klaster (c6) cine profesori sa akademije koji koriste i Python i R, ali cesce Python. Prilikom klasterovanja, atribut Current role ima najmanju ocenu znacajnosti, i ne ucestvuje u konstrukciji klastera.

3.3.2 Klasterovanje prema zaradama

Sledi prikaz klasterovanja koje grupiše ispitanike prema starosti, visini obrazovanja, broju godina iskustva i godišnjoj zaradi. Klasterovanje je izvršeno algoritmom K-Sredina, i rezultati su prikazani narednoj tabeli.

c0	18-21	BSc	0-1	0-10,000
c1	22-24	MSc	0-1	Didn't want to disclose
c2	30-34	Phd	5-10	Didn't want to disclose
c3	25-29	MSc	0-1	0-10,000
c4	22-24	MSc		

Tabela 5: Reprezentativne vrednosti klastera

Može se uočiti da veliki broj ispitanika čine studenti (ukupno 6883) sa iskustvom iz oblasti mašinskog učenja manjim od godinu dana (klaster c0). Takođe, klaster c4 čine pretežno doktori (ukupno 3992) sa iskustvom između 5-10 godina.

3.4 Klasterovanje Kohenen algoritmom

Kohenenov algoritam predstavlja posebnu vrstu neuronske mreže za učenje bez nadzora. Vrší prostorno mapiranje objekata u 2D prostor, pri čemu se slične instance grupiše. Kohenenova mreža ima dva sloja: ulazni i izlazni, koji su u potpunosti povezani. Ulazni čvorovi odgovaraju atributima, a izlazni klasterima.

3.4.1 Klasterovanje prema jezicima i visini obrazovanja

Sledi prikaz klasterovanja izvršeno Kohenenovim algoritmom. Jednostavan model mreže pronalazi 12 klastera i ima senka koeficijent 0.4. Rezultat je prikazan narednom tabelom.

c0	Python	Phd	Academia
c1	Python	BSc	Academia
c2	Python	BSc	Student
c3	Python	Phd	Comp/Tech
c4	R	Prof	Student
c5	PythonR	BSc	Comp/Tech
c6	Python	MSc	Comp/Tech
c7	R	MSc	Student
c8	PythonR	Phd	Academia
c9	Python	MSc	Student
c10	R	MSc	Comp/Tech
c11	PythonR	MSc	Comp/Tech

Tabela 6: Reprezentativne vrednosti klastera

Može se uočiti nekoliko klastera. Klaster c0 koji čine pretežno doktori, ima ukupno 1557 instanci (12.3%), koji rade na akademiji i imaju između 1 i 2 godine iskustva. Zatim klaster c8 koji ima ukupno 959 (7.6%) instanci,

koji čine doktori između 5-10 godina iskustva. Dva vrlo slična klastera c2 i c9 sa potpuno istim vrednostima, koji uglavnom čine studenti. Imaju ukupno 2336 instanci, što čini 18.5 % ukpnog broja klasterovanih instanci.

4 Zaključak

Prilikom klasterovanja, najbolje se pokazao KModes algoritam korišćen u Python programskom jeziku. Ovaj algoritam ima bolje performanse i rezultat, za razliku od algoritama K-Sredina i Kohenen. Za razliku od njega, algoritam K-Sredina zahteva dodatnu obradu kategoričkih podataka.

5 Literatura

Literatura

- [1] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.