Analiza podataka o mašinskom učenju i istraživanju podataka

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Petar Mićić
micicpetar73@gmail.com

15. avgust 2019

Sažetak

ovde napisati apstrakt

Sadržaj

1 Uvod

Oktobra 2018. godine je postavljena anketa o mašinskom ucenju 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, Kohenen 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 ponudjeninh odgovora (eng. mutiple choice responses). Primeri pitanja iz ankete:

- odaberite programske jezike koje koristite redovno
- odaberite programske jezike koje biste preporučili da nauče istrazivači 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. Na slici 1 je prikazan deo podataka.

	sex	age	education level	salary	prog_lang_part_1	ml_lib_part_1	data_vis_part_1
1	Male	35-39	Master's degree		Python	Scikit-Learn	ggplot2
2	Male	25-29	Bachelor's degree	I do not wish to disclose my approximate yearly compensation	Python	Scikit-Learn	
3	Male	18-21	Master's degree	0-10,000	Python	Scikit-Learn	ggplot2
4	Male	30-34	Master's degree	20-30,000	Python	Scikit-Learn	ggplot2
5	Male	22-24	Bachelor's degree	I do not wish to disclose my approximate yearly compensation	Python	Scikit-Learn	ggplot2
6	Male	40-44	Master's degree	125-150,000	Python	Scikit-Learn	
7	Male	25-29	Doctoral degree	30-40,000	Python	Scikit-Learn	ggplot2
8	Male	25-29	Bachelor's degree	30-40,000	Python	Scikit-Learn	ggplot2
9	Female	25-29	Bachelor's degree	10-20,000	Python	Scikit-Learn	ggplot2
10	Male	25-29	Master's degree	30-40,000	Python	Scikit-Learn	
11	Male	40-44	Master's degree	50-60,000			ggplot2
12	Male	25-29	Professional degree	0-10,000	Python	Scikit-Learn	
13	Male	35-39	Some college/university study without earning a bachelor's degree	100-125,000	Python	Scikit-Learn	ggplot2
14	Male	55-59	Some college/university study without earning a bachelor's degree	30-40,000	Python	Scikit-Learn	
15	Male	30-34	Bachelor's degree	90-100,000	Python	Scikit-Learn	ggplot2
16	Female	25-29	Master's degree	0-10,000	Python	Scikit-Learn	
17	Male	30-34	Master's degree	70-80,000	Python	Scikit-Learn	ggplot2
18	Male	25-29	Master's degree	0-10,000	Python	Scikit-Learn	
19	Male	30-34	Bachelor's degree	0-10,000	Python	Scikit-Learn	ggplot2
20	Female	25-29	Bachelor's degree	10-20,000	Python	Scikit-Learn	ggplot2

Slika 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 ispitancima 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 2.

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 analizing 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 2: 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 medju podacima nalazi veliki broj kategoričkih atributa, biće izvršeno klasterovanje algoritmima: K-Modes, K-Means, Kohenen. Klasterovanje ce biti usmereno samo na instance koje imaju popunjena polja Python i R (dakle da li ispitanik koristi Python, R ili oba).

3.1 Problem kategoričkog klasterovanja

3.2 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, obradjeni 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.

3.3 Klasterovanje koriscenjem KModes algoritma

3.3.1 Primer

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 3.3.1

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

Reprezentativne vrednosti klastera

Kod klasterovanja ispitanika koji koriste Python, R ili oba programska jezika moze 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 znaci da većina ne koristi tu biblioteku.

3.3.2 Primer 2

Grupsanje 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						

3.4 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 nacin. 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 0((n+k)*d), pri čemu je n - broj tacaka, k - broj centroida, d - dimenzionalnost a i - broj iteracija.

3.4.1 Primer

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 izvrseno algoritmom K-Means, pri cemu 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

Mogu se uociti 3 klastera. Prvi klaster (c1) cine pretezno studenti koji najcesce koriste programski jezik Python i su zavrsili ili zavrsavaju osnovne studije. Drugi klaster (c5) cine pretezno 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.

3.4.2 Primer 2

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 u 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		

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

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 Minig: The Textbook*. Springer, 2015.