Univerzitet u Beogradu
Matematički fakultet

Seminarski rad
iz predmeta
Naučno izračunavanje

Tema:

Vrednosti van granica

Mentor:
Anđelka Zečević

Student:
Ana Vuksić 1086/2016

Septembar 2018

Sadržaj

Poglavlje 1

Uvod

U današnje vreme podaci igraju veliku ulogu u našim životima. Trenutno se dnevno proizvede oko 2.5 EB podataka, što u vidu slika i videa do objava na socijalnim mrežama. Smatra se da će se ovaj broj samo povećavati. Dobar deo podataka se obrađuje i koristi za dolaženje do skrivenih, nama interesantnih informacija, kao na primer da bi se zaštitili od mogućih napada na našu mrežu, pa i kako bi predvideli prirodne nepogode. Kako se može doći do ovih informacija korišćenjem dobijenih podatak? Tu nam veliku ulogu igra analiza podataka i pojavljivanje vrednosti van granica.

Šta su vrednosti van granica? Vrednosti van granica su podaci čija vrednost je značajno udaljena i odskače od drugih njemu bliskih vrednosti. Vrednosti van granica možemo posmatrati kao suprotnost od klasterovanja. Dok klasterovanje teži da sjedini u grupu slične podatke, ovi podaci predstavljaju individualne vrednosti koje se ne uklapaju u našu širu sliku. Još ih nazivamo i abnormalnostima, anomalijama, devijantama i slično.

Ali zašto je identifikacija potencijalnih vrednosti van granica nama bitna, čemu one služe?

- Vrednosti van granica mogu biti indikatori loših podataka. Na primer, podaci mogu biti kodirani nepravilno ili neki eksperiment može biti pogrešno izveden. Ako potvrdimo da je neka vrednost tu zbog greške, onda bi trebalo da je izbrišemo(ignorišemo) ili ako je moguće popravimo njenu vrednost.

- Nekad ti podaci ne predstavljaju greške. U tom slučaju kad ne predstavljaju greške postoji mogućnost da nam oni ukazuju na neke naučne zanimljivosti. Mogu nam ukazivati na nedoslednoti u podacima i time upozoriti na moguću opasnu situaciju (razni internet napadi i prevare).

Vrednosti van granica su oni podaci čije vrednosti se nužno ne uklapaju u normalni model podataka. Koliko je zapravo van granica nam može reći težina vrednosti van granica. Većina algoritama koja pronalazi vrednosti van granica računa i težinu kao povratnu vrednost. Može vratiti broj ili binarni rezultat koji bi prikazao ovu vrednost. Ako vrati kao binarni broj, onda imamo samo jedinicu ili nulu koji govore da li jeste ili nije vrednost van granice.

Sad kada znamo da oni mogu biti greške, ali i izuzeci ili nedoslednosti u podacima, kako da odlučimo da li su bitni ili ne? Pre nego što krenemo sa analizom i traženjem odgovora na to da li treba da ih ignorišemo, mi moramo da znamo kako da ih nađemo. Jedan način je pomoću vizuelizacije i to ćemo preći u sledećoj glavi.

Poglavlje 2

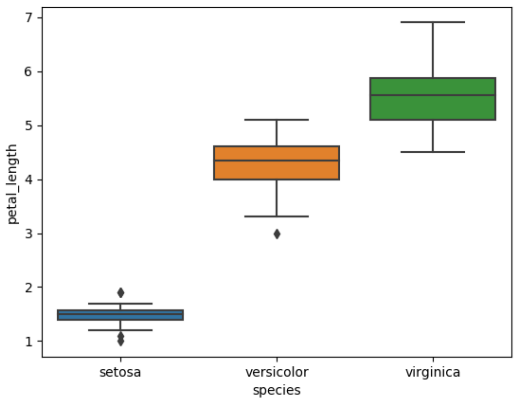
Vizuelna predstava

U ovom poglavlju ćemo preći razne grafove(vizuelne metode) pomoću kojih možemo lako doći do nekih zaključaka. Koristićemo graf sa kutijicama (eng. boxplot), histogram i raštrkani graf. Za prikaz svih grafika ćemo koristiti isti skup podataka koji se naziva Iris i ima sledeće podatke.

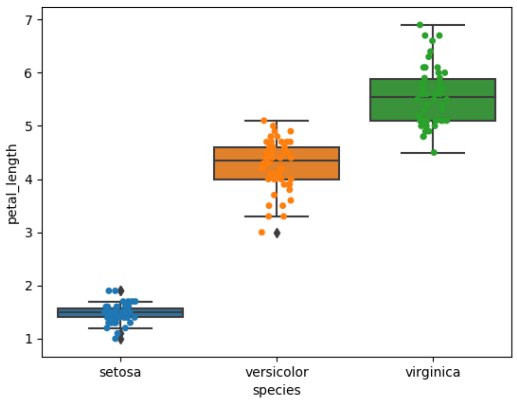
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Slika 1:Iris skup podataka

Pogledajmo prvo kako izgleda jedan grafik sa kutijicama. Posmatraćemo verziju gde su prikazane sve vrednosti i verziju bez prikaza svih vrednosti već samo oznaka za percentile.



Slika 2.1:Bez prikaza vrednosti

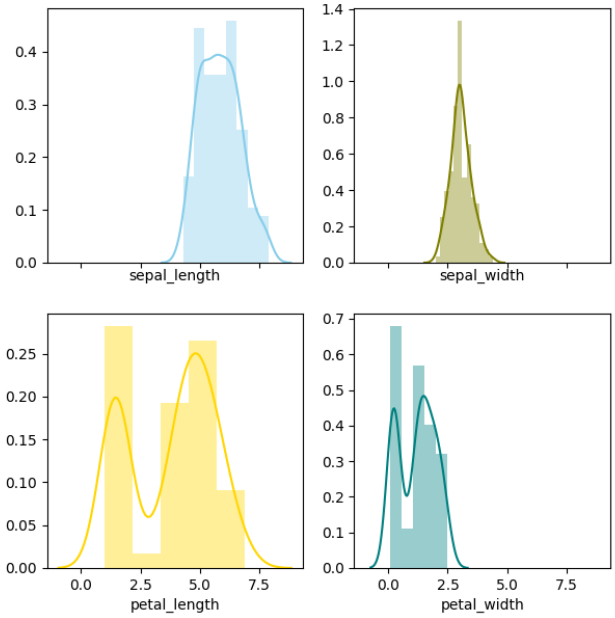


Slika 2.2:Sa prikazanim vrednostima

Na grafiku bez vrednosti možemo lako uvideti gde se nalaze vrednosti van granica. One se nalaze posle krajnjih linija.

Sledeći grafik koji ćemo posmatrati je histogram (Slika 3). Kod histograma imamo pravougaonike koji nam govore o tome kolika je vrednost podataka u skupu po određenim atributima. Kada imamo neke pravougaonike koji su izolovani to nam govori da su to verovatno vrednosti van

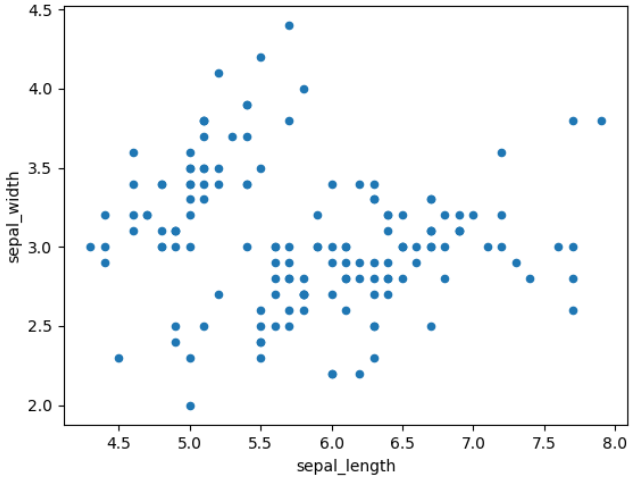
granica.



Slika 3: Histogram

Grafik sa raštrkanim vrednostima se nalazi na slici ispod(Slika 4). Ovde su sve klase u podacima Iris predstavljene odjednom. Kada je neka tačka jako izolovana to nam govori da je u pitanju vrednost van granica.

Često ljudi misle da će sam pogled na podatke da nam bude dovoljan da nađemo ove vrednosti. Ovo možda i jeste tačno na malim primerima, ali zamislimo da imamo podatke sa više od 600 kolona i 10 000 redova. Da li bi i dalje pogled na podatke bio dovoljan? Ovo nam govori da se ne možemo osloniti na detekciju ovih podataka samo pomoću grafova. Imamo dosta statističkih metoda koje nam tu mogu pomoći i drugih raznih modela.



Slika 4:Raštrkane vrednosti

Poglavlje 3

Analiza vrednosti van granica

U narednom delu ćemo pomenuti neke od najkorišćenijih metoda za analizu vrednosti van granica. Kasnije ćemo neke od njih opisati detaljnije.

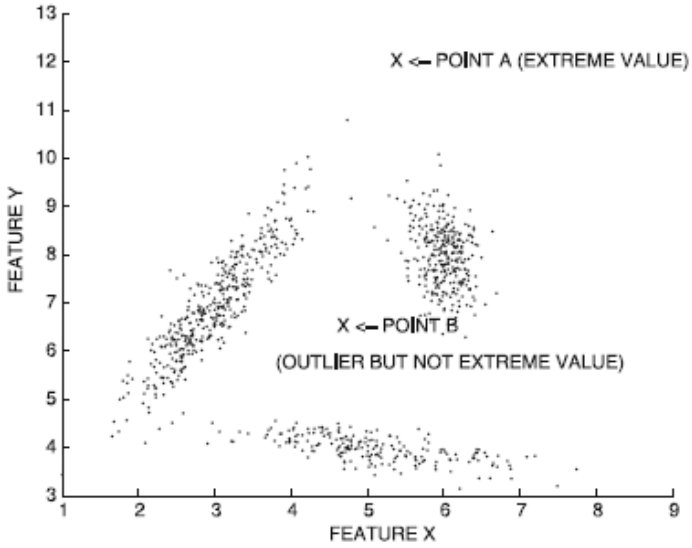
- Ekstremne vrednosti: Podatak je ekstremna vrednost, ako se nalazi na jednoj od dve granice verovatnosne raspodele. Možemo definisati ekstremne vrednosti i kod multidimenzionanih podataka.
- Klasterovanje: Ono se smatra suprotnim problemom od detekcije vrednosti van granica. Jedan traži podatke koji se nalaze zajedno u grupama, a drugi traži izolovane podatke. Neki algoritmi za klasterovanje sami propratno nalaze vrednosti van granica. Ili ih možemo modifikovati tako da baš traže ove vrednosti.
- Modeli zasnovani na udaljenosti: U ovom slučaju algoritam k najbližih suseda se analizira kako bi našli vrednosti van granica. Intuitivno ako je vrednost k najbližih suseda velika to nam govori da tačka može biti vrednost van granica.
- Modeli zasnovani na gustini: Kod njih se koristi gustina lokalnih podataka kako bi utvdili težinu vrednosti van granice. Modeli koji su zasnovani na gustini su jako bliski sa algoritmima koji su zasnovani na udaljenosti. Kada je gustina niska, to znači da je udaljenost među podacima veća.
- Verovatnosni modeli: Ovo je jedan način za nalaženje klastera. Kao što smo već rekli klasterovanje i detekcija vrednosti van granica su komplementarni problemi. Koraci su jako slični sa koracima pri korišćenju modela klasterovanja. Osim što je EM algoritam korišćen za klasterovanje, i verovatnosno podobne vrednosti su korišćene za računanje težina vrednosti van granica, umesto udaljenosti.
- Informaciono-teoretski model: Ovi metodi imaju zanimljivu povezanost sa prethodnim modelima. Većina modela posmatra udaljenost mogućih vrednosti van granica od ostatka modela i tako ih nalazi. Ovi metodi izbacuju one sa maksimalnom devijacijom i vide promene u rezultatima. Ako je razlika velika, oni su vrednosti van granica.

3.1 Analiza ekstremnih vrednosti

Podaci koji se nalze na ivicama skupa podataka se označavaju kao podaci van granica. Oni odgovaraju repovima pri raspodelama verovatnoće. Iako se lakše u statistici prikazuju repovi za 1-dimenzionalne raspodele, treba da znamo da se one mogu prikazati i za multidimenzionalne

podatake. Treba imati na umu da su ekstremne vrednosti specijalni slučajevi vrednosti van granica. Tj. sve ekstremne vrednosti su vrednosti van granica, ali obrnuto ne mora da važi.

Primer 1.1: Posmatrajmo 1-dimenzionalni skup podataka $S = \{1, 3, 3, 3, 50, 97, 97, 97, 100\}$. Vrednosti 1 i 100 možemo smatrati ekstremnim vrednostima. Vrednost 50 je srednja vrednost skupa podataka i nije ekstremna vrednost. Posmatranjem ovog skupa možemo da primetimo da to i jeste najizolovanija tačka među podacima i iz tog ugla posmatranja treba da bude smatrana vrednošću van granica, ali nije. Slično važi i za multidimenzionalne podatke, iako je teže definisati rep kod ovih raspodela.



Slika 5: Mutidimenzionalni podaci

Posmatrajmo sliku 5. Imamo tačku A koja je i ekstremba vrednost i vrednost van granica, dok je tačka B samo vrednost van granica ali nije ekstremna.

Bitnosti ove metode je da on prebacuje težinu i verovatnoću da je nešto vrednost van granica u binarnu labelu tako što identifikuje one koji su ekstremne vrednosti.
Neka pitanja koja bi trebali sebi da postavimo kod ovih analiza jesu sledeća.

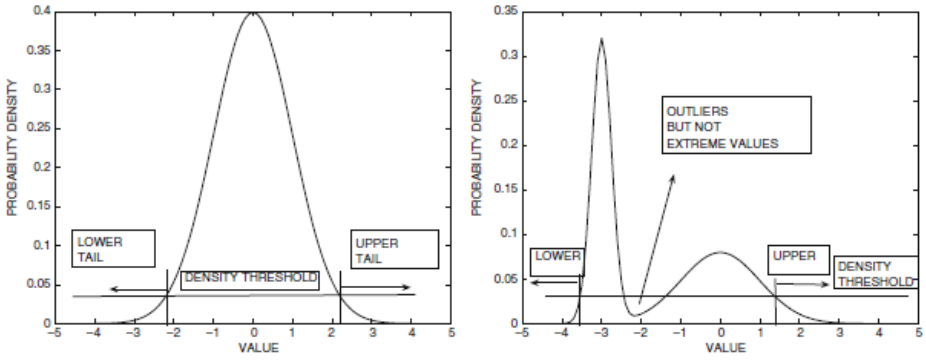
- Da li su u pitanju jednodimenzionalni ili više dimenzionalni podaci?
- Da li možemo pretpostaviti raspodelu podataka.

3.1.1 Analiza ekstremnih vrednosti jednodimenzionalnih podataka

Ovo je usko povezano sa statistikom. Oslanjamo se na razna izračunavanja pri raspodelama. Gledamo repove raspodela i proveravamo podatke koji se tu nalaze. Uglavnom se smatra da podaci već imaju određenu raspodelu. Ovi metodi pokušavaju odrede deo podataka od kojih se očekuje da budu ekstremne, na osnovu ovih raspodela. Ove vrednosti nam govore sa kojim uverenjem možemo tvrditi da je neka tačka ekstremna vrednost.

Kako zapravo definišemo *rep podataka*? Za nesimetrične raspodele, smisleno je pričati o gornjem i donjem repu, koji možda neće imati istu verovatnoću.

Gornji rep su sve ekstremne vrednosti veće od određenje vrednosti, a donji rep sve manje od određene vrednosti. Na slici 6 su prikazani repovi za simetričnu i nesimetričnu podelu.



Slika 6: Repovi raspodela

Tu lako uviđamo da oblast u gornjem i donjem repu nesimetrične podele ne mora biti ista. Iako ima i u oblasti ovog grafa dodatnih mesta sa manjom gustinom ne smatramo ih ekstremnim vrednostima jer nisu u repu. Mogu biti smatrani vrednostima van granice ali ne ekstremnim vrednostima.

U simetričnim verovatnosnim raspodelama, rep je definisan u okviru ovih regija, pre nego preko gustine. Ipak gustina ostaje kao karakteristika repa, posebno kod asimetričnih jednodimenzionalnih podataka. Neke asimetrične raspodele možda čak i neće imati rep na jednom kraju (eksponencijalna).

Najkorišćeniji model je model za *normalnu raspodelu*. On ima funkciju gustine fX(x) sa srednjom vrednošću μ i standardnom devijacijom σ.

Kod normalne raspodele je srednja vrednost 0 a standardna devijacija 1. Ove vrednosti možemo odrediti i iz samog skupa podataka sa velikom tačnošću. Možemo ih iskoristiti kako bi došli do vrednosti *Z testa*.

$$Z = (X - \mu)/\sigma \tag{3.1}$$

Velike pozitivne Z vrednosti odgovaraju gornjem repu, dok su velike negativne vrednosti odgovara donjem repu. Na gornjoj slici možemo uvideti da će vrednosti veće od 3 biti smatrane ekstremnim vrednostima, to je Z broj 3. Oblast će biti manja od 0.01

Ako je broj podataka manji može se koristiti studentska raspodela umesto normalne. Kad je n veliko, veliki broj podataka, ona konvergira ka normalnoj raspodeli.

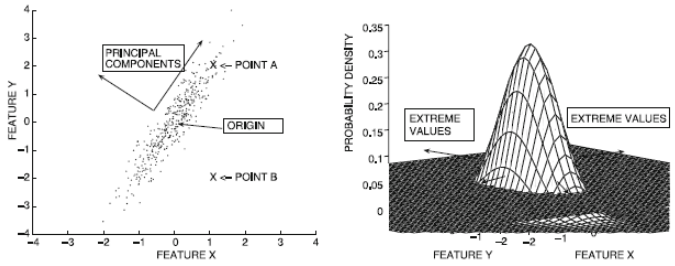
3.1.2 Analiza multidimenzionalnih ekstremnih vrednosti

Kao što imamo definisane repove za jednodimenzionalne podatke, kao region sa manjom gustinom od određene gustine, slične metode možemo koristiti i na višedimenzionalnim podacima. Ovde se pojavljuju verovatnosne podele sa jednim šiljkom.

Možemo koristiti Gausovski model za višedimenzonalne podatke. Pretpodstavljamo da su svi podaci u verovatnosnoj podeli sa jednim šiljkom (eng. single Gaussian cluster), i podaci najudaljeni od ovoga se smatraju ekstremnim vrednostima.

Možemo koristiti razne formule za računanje udaljenosti (nećemo se na njih osvrtati u ovom radu). Jedno od mogućih je i pomoću Mahalanobisovog rastojanja između srednje vrednosti i tog podatka. Da bi verovatnoća gustine pala ispod određene vrednosti, Mahalanobisovo rastojanje treba da bude veće od određene vrednosti. Tako ovu udaljenost možemo koristiti za označavanje težine ekstremne vrednosti. Veća vrednost bi pokazivala na veću mogućnost da je podatak ekstremna vrednost.

$$Maha(X,Y) = \sqrt{(X-Y)\sum(X-Y)^T} \tag{3.2}$$



Slika 7: Multidimenzionalni podaci

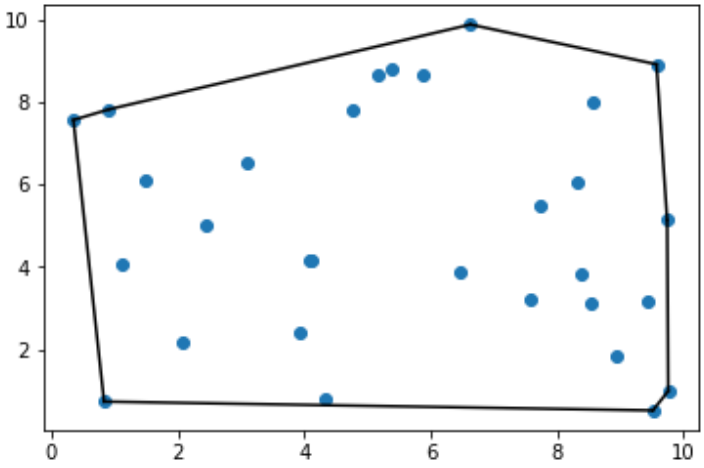
3.1.3 Metode zasnovane na dubini

Dobar predstavnik ovih metoda je zasnovan na konveksnom omotaču. Konveksni omotači predstaljaju skup sa mogućim ekstremnim vrednostima. To može biti iterativni algoritam, gde u k-tom koraku, sklanjamo sve tačke iz skupa koji su konveksni omotač u tom koraku. U ovim koracima možemo da damo i k kao težinu vrednosti van granica i što je manja vrednost broja k to je veća tendencija da je vrednost van granice. Radimo ovo dok se skup ne isprazni. Možemo preobratiti i u binarnu labelu gde stavimo da je onaj sa najmanjim k-a jedan. Ovde je prikaz algoritma za nalaženja konveksnog omotača, možemo pomoću izbacivanja iskorišćenih tačaka da nađemo sve konveksne omotače.

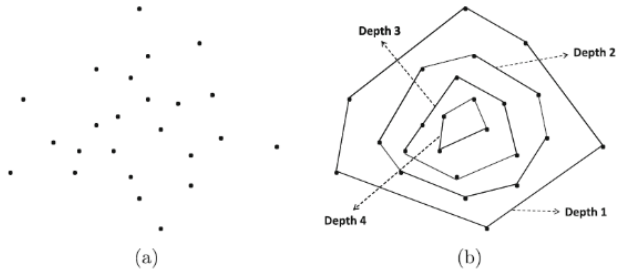
```
from scipy.spatial import ConvexHull
import numpy as np
import matplotlib.pyplot as plt

points = np.random.rand(30, 2)   # 30 random points
hull = ConvexHull(points)

plt.plot(points[:,0], points[:,1], 'o')
for simplex in hull.simplices:
     plt.plot(points[simplex, 0], points[simplex, 1], 'k-')

plt.show()
```

Slika prikazuje konveksni omotač sa ekstremnim vrednostima. Možemo povezati ovaj algoritam sa slojevima na luku i njihovim skidanjem. Svi podaci u omotaču se tretiraju isto, ovo je nepoželjno.

Osim toga, rasprostranjenost tačaka na uglovima konveksnog omotača uglavnom se povecava sa dimenzionalnošcu. Za veoma visoku dimenzionalnost dešava se da se većina podataka nalazi u uglovima prvog omotača. Kao rezultat, ne moguce je razlikovati težinu različitih podataka. Računarska složenost ovog metoda značajno se povecava sa dimenzionalnošcu. Kombinacija kvalitativnih i računskih problema povezanih sa ovom metodom čini je lošom.



Slika 8: Konveksni omotač, ekstremne vrednosti



Slika 9: Konveksni omotač kroz iteracije

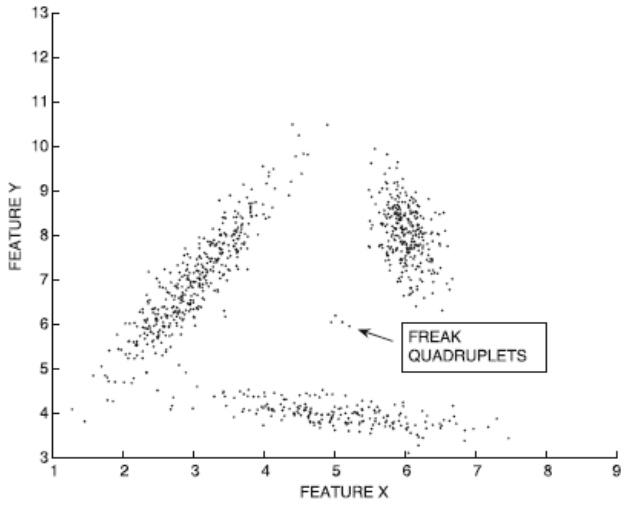
3.2 Analiza pomoću klasterovanja

Već smo ranije pominjali klastrerovanje i njegovu poveznost sa detekcijom vrednosti van granica. Klasterovanje nam služi za pronalazak grupa podataka, dok je analiza vrednosti van granica pomoću klasterovanja tu da nadje vrednosti koje se ne uklapaju u nađene grupe. Štaviše ove tačke što su dalje od tih grupa je jasnije da su u pitanju vrednosti van granica. Tako da vrednosti van granica i klasterovanje imaju komplementarnu vezu. možemo reći da je tačka ili deo klastera (grupe) ili vrednost van granica.

Algoritmi za klasterovanje često imaju opcije za rad sa vrednostima van granica u vidu uklanjanja istih. Detekcija vrednosti van granica pomoću klastera nije preporučljiva. Razlog za to bi bio jer oni nisu optimizovani za ovo. Vrednosti koje se nalaze na granicama klastera se mogu smatrati blagim vrednostima van granica, ali one kao tako označene su često niodkakvog značaja.

Ovi metodi imaju neke i prednosti. Vrednosti van granica često imaju tendenciju da se pojave u obliku svog manjeg klastera. Anomalija zbog koje je su nastali se može pojaviti više puta na više njih. Kao rezultat tog ponavljanja javlja se mala grupa vrednosti van granica. Primer ovoga se nalazi na slici 10 .

Težinu vrednosti van granica nije toliko teško naći, to bi nam bila udaljenost tačke od centroida klastera. Jedna od mera koje je dobro koristiti za ovo je Mahalanobisovo rastojanje. U delu sa multidimenzionalnim ekstremnim vrednostima smo pričali o ovome. Lokalno Maha-

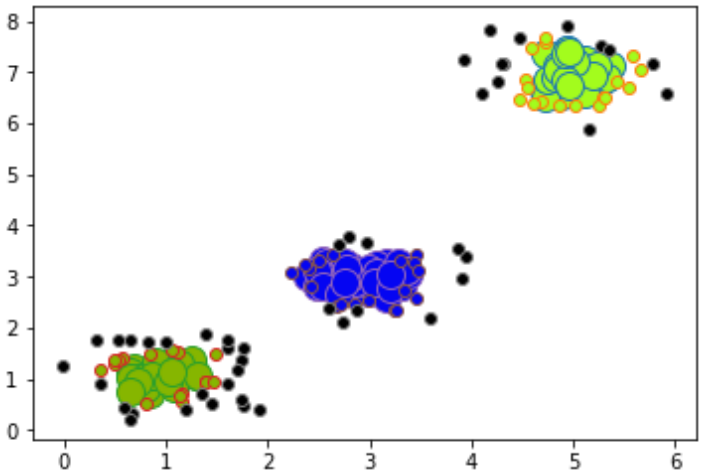


Slika 10: Klasterovanje sa vrednostima van granica

lanobisovo rastojanje je više značajno za otkrivanje generalno vrednosti van granica, dok je globalno Mahalanobisovo rastojanje više za vrednosti van granica koje su i ekstremne vrednosti. Algoritmi za klasterovanje su većinski pravljeni za globalne podatke. Što bi značilo da neku malu grupu od recimo 4 tačke često i neće obeležiti kao klaster iako će imati veliku težinu vrednosti van granice. To je iz razloga što većina algoritama traži neki minimalni broj tačaka, neku određenu masu. Što i nije slučaj sa algoritmima koji rade na osnovu gustine, pogotovo ako je koncentracija na lokalu.

Glavni problem sa algoritmima za klasterovanje je što nekad ne može da napravi razliku između šuma i prave anomalije. Ovi podaci neće završiti u klasteru i njihova udaljenost od centroida neće predstavljati pravu težinu vrednosti van granica. Problem su i male grupe, koje neće detektovati.

Imamo grafički prikaz klasterovanja korišćenjem algoritma DBSCAN.



Slika 11: DBSCAN, crnu su vrednosti van granica

3.3 Analiza pomoću modela baziranih na gustini

Metodi bazirani na gustini su bliski algoritmima za klasterovanje (recimo DBSCAN) koji su bazirani na gustini. Ideja je da se determinišu regioni u podacima kako bi mogli da nađemo vrednosti van granica.

Mogu se koristiti metodi zasnovani na histogramima, rešetkama ili kernelima. Histograme možemo posmatrati kao specijalan slučaj jednodimenzionalne rešetke. Ovi metodi nisu uspeli da dostignu neku veću popularnost, zbog poteškoća da prilagode gustinu u određenim delovima. Takođe i što je veća dimenzionalnost teže je odrediti gustinu. Više se koriste kod jednodimenzionalnih podataka.

3.3.1 Baziran na histogramu i rešetkama

Histograme je za male dimenzije podataka lako napraviti i intuitivni su pa ih često koristimo. u ovom slučaju podaci su diskretizovani u korpe i frekvencija svake korpe je određena. Oni podaci koji se nalaze u korpama sa niskom frekvencijom nazivamo vrednostima van granica.

Kod višediezionalnih podataka, koristimo strukturu rešetke. Svaka dimenzija je particionisana u p jednakih delova. Tačke podataka sa gustinom manjom od određene spada u vrednosti van granica.

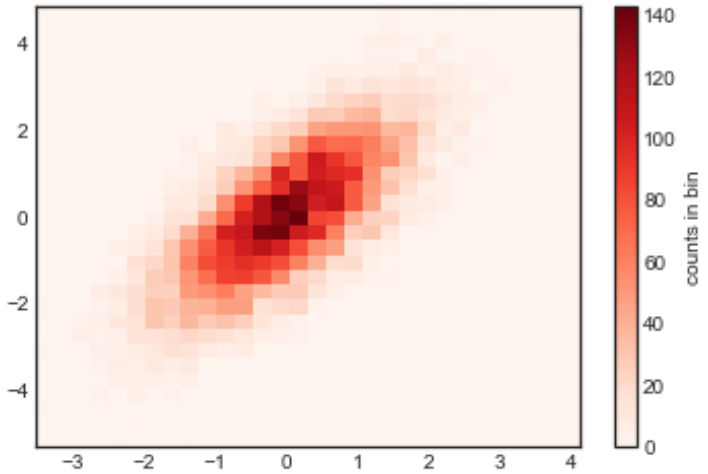
Jedan od problema je naći dobru širinu korpi/rešetke. Ako su preširoki ili preuski neće dobro modelirati raspodelu frekvencije. Preširoko, dolazi do prihvatanja i vrednosti van granica, preusko i oni su klasifikovani kao vrednosti van granica.

Drugi problem je to što ovi algoritmi često gledaju samo manji lokalni deo podataka, a ne globalne karakteristike podataka.

Na primer u slici iz dela sa klasterima, ne bih mogao da klasifikuje one podatke kao vrednosti van granica. Osim u slučaju baš dobrog baratanja rešetkom. problem je i kada su velika odstupanja u gustini među rešetkom.

Histogrami su komplikovani kada su u pitanju višedimenzionalni podaci. Alternativa je prvi naći manju dimenzionalnost tih podataka. Ovi problemi su dobro poznati, a znaju da se pojave i kod drugih metoda kao što je klasterovanje.

Na slici možemo videti kako može izgledati podela na rešetke i prebrojavanje broja pojavljivanja podataka u određenom polju. Koristili smo hist2d funkciju iz matplotlib biblioteke.



Slika 12.1: Resetke

3.3.2 Kernel

Nećemo zalaziti preterano u dubinu ovog metoda. Napomenućemo da je ovo dobar način da "razmažemo" tačke i dobijemo glatkiju funkciju. Ove metode su veoma slične metodama sa histogramima. Razlika je što pomoću ove metode dobijamo ravniji profil gustine.

U proceni gustine pomoću kernela, neprekidna procena gustine je generisan u svakom momentu. Vrednost gustine nekog dela je procenjena kao suma glatkih delova kernel funkcije $Kh()$. Svaki kernel funkcija je povezana sa kernelom širine h koja determiniše glatkoću.

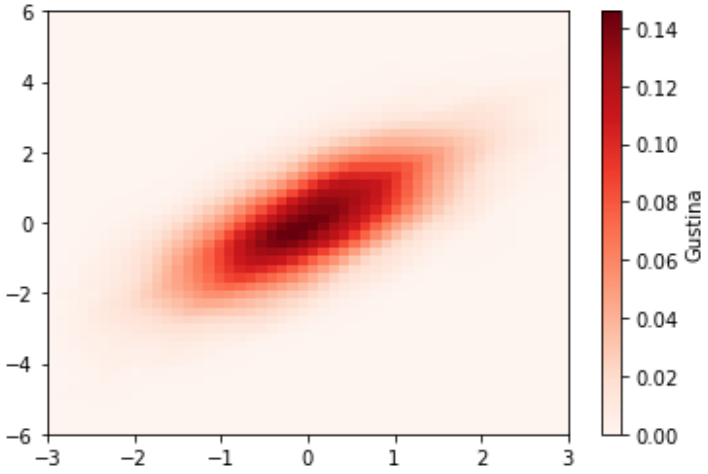
Svaka vrednost Xi je zamenjena sa neprekidnom funkcijom $Kh()$ čiji je vrh, najveća vrednost, u Xi i ima varijansu određenu od strane parametra h.

Primer Gausovskog kernela sa širinom h.

$$Kh(X - Xi) = (1/\sqrt{2\pi} * h)^d * e^{||XXi||2/(2h2)} \tag{3.3}$$

Greška je definisana preko širine h, koja je odeđeno pomoću podataka. Gustina je određena bez korišćenja date tačke. Vrednost gustine je težina vrednosti van granica. Što je gustina niža to je veća verovatnoća da je vrednost van granica u pitanju. Slične su poteškoće kao i kod metoda sa histogramima i rešetkama.

Možemo imati problem ako koristimo globalnog h, a da imamo velike varijacije u gustini. Mogu se poboljšati rezultati ako se orjentišemo više lokalno. Nisu baš efikasne za višedimenzionalne podatke. Tačnost pretpostavke gustine je lošija sa većom dmenzionalnošću.



Slika 12.2: Korišćenjem kernela

Poglavlje 4

Detekcija vrednosti van granica kod kategoričkih podataka

U ovoj oblasti nam je od velike važnosti tip podataka nad kojim treba da koristimo određene algoritme. Od tipa nam mogu zavisiti i način na koji koristimo, a i algoritam koji koristimo. Zato ćemo ukratko prokomentarisati moguće izmene kako bi analizirali vrednosti van granica kod kategoričkih podataka. Ovde su potrebne manje promene nego možda za neke druge oblasti u istraživanju podataka, kao na primer klasterovanje. U nastavku ćemo pričati o nekim od mogućih promena nad metodama.

4.0.1 Metode zasnovane na klasterovanju i udaljenosti

Kod ovih metoda imamo dve glavne izmene:

- Potrebno je koristiti metode za klasterovanje koje su dobre nad kategoričkim podacima. Lako možemo od kategoričkij da dođemo do binarnih podatak i da onda radimo na njima.
- Bitan je odabir funkcije sličnosti između ovih podataka.

4.0.2 Biarni i podaci iz određenog skupa

Ovo je specijalan vid kategoričkih podataka. koji se često javlja. Štaviše, i kategorički i numerički podaci uvek mogu da se predstave u ovom binarnom obliku. Analiza čestih stavki ili algoritmi pravile pridruživanja su podrutina za detekciju vrednosti van granica. Glavna ideja je da se vrednosti van granica ne pojavljuju u čestim stavkama. Tako da je jedan od mogućih rešenja da se koristi suma svih podrški čestih stavki u određenoj transakciji. Suma bi bila normalizovana podelom sa brojem čestih stavki.

Intuintivno, transakcija sa velikim brojem čestih stavki sa velikom podrškom će imati imaće visoku vrednost. Za takvu transakcija je mala verovatnoća da će biti vrednost van granica, jer pokazuje na pravilo među podacima. Tako da se one sa manjom vrednošću pre smatraju vrednostima van granica. Problem kod ovog pristupa bi bio, da možda ne bi mogao da napravi razliku između stvarno izolovanih tačaka i šumova. Jer nijedna od ovih vrednosti se neće često pojavljivati. tako da će biti isto čitane. Iz ovog razloga, ovaj postupak nas ne dovodi uvek do najvećih anomalija među podacima.

Poglavlje 5

Moguća rešenja za vrednosti van granica

Kada nađemo vrednosti van granica šta sa njima možemo da radimo? Sve naravno zavisi od samog konteksta. Moramo biti dobro upoznati sa podacima kako bi znali da li su nam te vrednosti bitne ili ne. Sad ćemo preći preko nekih mogućih rešenja za rad sa vrenostima van granica.

Odstranjivanje. Ovo je jedan od najlakših i najbržih načina za obradu vrednosti van granica.Sve tačke koje su vrednosti van granica uzimamo i ako ne utiču značajno na vrednost ostatka skupa podataka, izbacujemo ih iz skupa.

Izbacivanje i čuvanje pojavljivanja. Obljasnićemo pomoću sledećeg primera. Zamislimo da imamo ljude koji skaču u dalj. Došao je novi trener i proveravamo da li je došlo do poboljšanja rezultata skakača. Rezultati su da su osobe redom imale skok +12,+7,+10,-56 cm veći ili manji nego pre. Ako nađemo srednju vrednost jer hoćemo da vidimo da li je bolje sa trenerom mi dobijemo srednju vrednost koja je negativna, ali ako izbacimo -56 srednja vrednost će biti pozitivna i veća. U ovom slučaju možemo odbaciti podatak koji nam je van vrednosti, ali bi trebalo da ga sačuvamo negde i imamo na umu.

Smena. Zamislimo da imamo vrednosti van granica u box plotu (slika 1.1), granice koje su tu prikazane kao granice i koje predstavljaju 25 i 75 percentil nisu minimum i maksimum u skupu. Mi možemo da uzmemo i zamenimo vrednosti van granica sa vrednošću ovih vrednosti. Ovako ćemo sačuvati neki uticaj koji mogu imati, a neće biti tako daleki podaci.

Transformacija promenljivih. Recimo možemo iskoristiti funkciju logaritma na podacima i time izgubiti vrednosti van granica jer ova funkcija po prirodi teži da skupi podatke.

Poglavlje 6

Primena

Primena detekcije vrednosti van granica je jako široka. Zalazi u mnoge domene. Neki od njih su detekcija prevara, razni napadi, finansijske prevare, analitika WEB logova. Mnoge od ovih delatnosti se sastoje od kompleksnih podataka i ne mogu biti rešeni samo sa ovim jednostavnijim algoritmima, moramo se dovijati na razne načine kao što je prikazano kod multidimenzionalnih podataka.

6.0.1 Kontrola kvaliteta i detekcija prevare

Dosta primena detekcije vrednosti van granica se nalazi u delatosti kontrole kvaliteta i detekcije greške. Nekada je dovoljno korišćenje analize ekstremnih vrednosti podataka sa jednom dimenzijom, a nekada su potrebne kompleksnije metode. Na primer, greške pri radu mašina u proizvodnji možemo računati brojem proizvoda napravljenih sa greškom u danu. Kada ima previše robe sa greškom to nam može reći da je greška u mašini. Ekstremne vrednosti, jednodimenzionalne su ovde dovoljne. Možemo pratiti razne delove mašine, njen motor recimo. Broj obrtaja, zagrevanje. Želimo da nađemo ovakve kvarove što pre. Ovo su uglavnom privremene analize i parametre moramo privići na njih.

6.0.2 Finansijska prevara

Ovo je isto jedna od čestih primena nalize vrednosti van granica.

Takve vrednosti se mogu javiti kod prevara sa kreditnim karticama, transakcijama osiguranja i razmena unutrašnjih informacija. Kompanija čuva transakcije svih korisnika. Svaka ima podatke o korisniku pomoću kojih ih možemo identifikovati. Sve transakcije koje se ne uklapaju u redovne transakcje ovog korisnika su moguće prevare. Na primer ako neko odjednom podiže veliku svotu novca ili na geografski čudnim lokacijama za ovog korisnika. Slično bi se posmatralo i za prevare sa osiguranjem. Što više vremenskih i geografskih podataka imamo to nam je ovakva detekcija lakša.

6.0.3 Analiza WEB poseta

Ponašanje korisnika na internetu se često prati na razne načine i to automatski. Anomalije u korišćenju ovih sajtova se lako mogu uvideti iz njihovih zapisa. Na primer, želimo da uđemo u sistem koji je zaštićen nekom šifrom. Niz akcija koje korisnik sprovodi su neuobičajene u poređenju sa ponašanjem drugih korisnika koji znaju svoju šifru. Najefikasniji model koji bi nam ovde koristio je optimizovan model za sekvencijalne podatke. Alternativno možemo prebaciti ovo u multidimenzionalni zapis podataka i onda raditi na njima.

6.0.4 Neodobreni upadi

Napadi u vidu bilo kojih neovlašćenih pokušaja da se pristupi podacima kojim ta osoba ne treba da ima pristup. Česti scenario je napad na hosta i napad na mrežu. Pri napadu na hosta, operativni sistem i njegovi zapisi su analizirani kako bi se našao problem. Ovi podaci nisu mnogo drugčiji od WEB zapisa. Pri napadu na mrežu, trenutne veze između podataka su slabije i oni mogu da budu posmatrani kao niz multidimenzinalnih podataka. Oni zahtevaju analizu vrednosti van granica za niz.

6.0.5 Primena u zemljanim naukama

Ova analiza je dobra za praćenje anomalija u zemlji i okruženju. Praćenje raznih varijacija u temperaturi. Pomoću njih možemo doći do klimatskih neočekivanih i loših promena na zemlji. Pa i predviđanje uragana.

Poglavlje 7

Zaključak

Problem analize i detekcije vrednosti van granica je važan jer ga je moguće primeniti na dosta problema današnjice. Česte metode za njihovu analizu su verovatnosne metode, klasterovanje, metodi bazirani na distanci, metodi bazirani na gustini, informaciono - teoretski metodi. Među svim ovim najpopularniji su metodi zasnovani na udaljenosti, ali su oni kompjuterski skupi.

Pored poznavanja metoda koje možemo koristiti jako nam je bitno i razumevanje podataka. Tako da je potrebno odvojiti i određeno vreme na njihovo razumevanje.
Sem što detektovanje vrednosti van granica može biti težak problem i sama validacija vrednosti van granica je težak problem. Zbog nenadgledanog učenja i takve prirode korišćenih algoritama. Uglavnom koristimo eksternu validaciju. Iako se pojavljuje dosta poteškoća pri radu na ovome, potrebno je dalje istraživati i biti istrajan. Analiza ovih podataka može upozoriti ljude i na moguće prirodne nepogode i time im spasiti život.

Poglavlje 8

Literatura

Data Mining, Aggarwal, Charu C.:
https://www.springer.com/gp/book/9783319141411

Python Documentation:
https://docs.python.org/3/

Engineering Statistics Book :
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

Mahalanobis distance :
https://en.wikipedia.org/wiki/Mahalanobis$_d istance$

Outliers:
http://www.statisticssolutions.com/univariate-and-multivariate-outliers/