Izveštaj
Analiza Parkinson 2 baze podataka

Katarina Mladenović , BI-16/2015, kaca96mladenovic@gmail.com

I. BAZA PODATAKA

Bazu podataka čine atributi izdvojeni iz 195 snimaka fonacija vokala izgovorenih od strane 31 pacijenta, od kojih je 23 dijagnostifikovano sa Parkinsonovom bolešću. U proseku je snimljeno po 6 fonacija za svakog pacijenta, u trajanju od 1 do 36 sekundi. Za svaki snimak je izračunato po 22 obeležja (baza podataka je dimenzija 195x22). Svi atributi su numerički.

Izdvojene atribute čine karakteristike snimljenih fonacija - *osnovna frekvencija, apsolutni zvučni pritisak,* različito računati *jitter* (varijacija osnovne frekvencije tokom ciklusa govora) i *shimmer* (varijacija amplitude govora tokom ciklusa), *odnos šuma i harmonika* (aplituda šuma relativna sa amplitudama tonalnih komponenti govora), *odnos harmonika i šuma,* kao i 4 mere za analizu nelinearnih vremenskih nizova.

Poremećaji u glasu su jedan od ranih i pouzdanih ukazatelja na Parkinsonovu bolest. Cilj analize atributa je pronalazak mere diskriminacije obolelih od zdravih pacijenata, kao i mere procene stanja bolesti pomoću snimanja glasa pacijenata. Lečenje ove bolesti je skupo i terapija nije uvek dostupna, pa se ovi podaci mogu iskoristiti u telemedicini kako bi se lečenjem na daljinu troškovi smanjili a pristupačnost terapiji povećala.

II. ANALIZA PODATAKA

Podaci su podeljeni u dve klase – u atribute koji pripadaju obolelim i zdravim pacijentima. Sve vrednosti su računate posebno za obe grupe i razlika rezultata treba da bude osnova za diferenciranje pacijenata. Urađeno je sledeće.

A. Dinamički i interkvartilni opseg

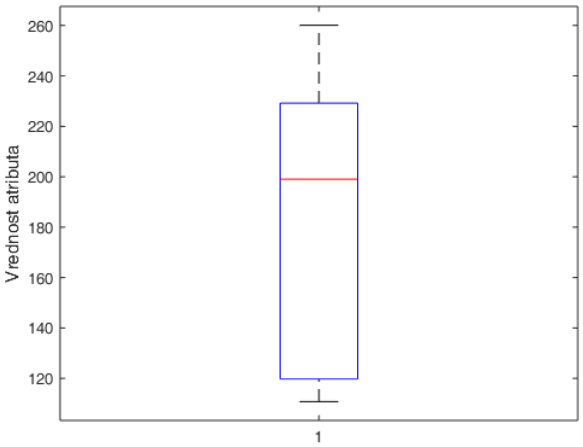
Dinamički opseg je dobijen kao razlika maksimuma i minimuma vrednosti podataka. Interkvartilni opseg je računat kao opseg u kome se nalazi 50% vrednosti podataka oko srednje vrednosti. Donja granica opsega je vrednost ispod koje se nalazi 25% podataka sa najnižim vrednostima, dok je gornja granica vrednost iznad koje se nalazi 25% podataka sa najvišim vrednostima. Atributi su različite prirode pa su i vrednosti dobijenih opsega raznolike.

B. Vrednost koja se javlja u više od 10% slučajeva

Traženo je postojanje ovakve vrednosti. Prvo je urađen histogram za svaki atribut sa binovima koji obuhvataju uzak opseg vrednosti, zatim je provereno da li postoji bin gde se nalazi više od 10% uzoraka. Ovakva vrednost je postojala kod svih atributa klase I, i kod 17 od 22 atributa klase II. Razlika u prisustvu može poslužiti za razdvajanje klasa.

C. Crtanje boxplot–a i prisustva outlier–a

Za svaki atribut je iscrtan boxplot. Zatim je utvrđeno prisustvo outliera – vrednosti koje izlaze van interkvartilnog opsega. Primer boxplot – a prikazan je na slici 1. Crvena linija je median atributa, plavi pravougaonik označava interkvartilni opseg a crvene zvezdice su outlier-i. Njihovo prisustvo je utvrđeno za sve atribute i izdvojeno kao poseban podatak. Zbog različite prirode atributa, boxplot–ovi su bili raznoliki. Outlier-e nisu imala samo 3 atributa klase I.

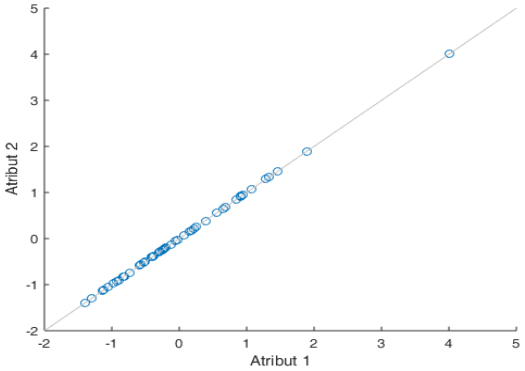


Sl. 1. Boxplot prvog atributa obolelih pacijenata.

D. Maksimalna korelacija atributa i analiza scatterplot–a

Za ovu analizu korišćeni su podaci nad kojima je izvršena Z normalizacija. Izračunata je korelacija svakog atributa sa svakim i pronađena maksimalna vrednost. Tako su dobijeni atributi koji su najviše korelisani (uticaj autokorelacije je uklonjen). Zatim je nacrtan scatterplot za ta dva atributa i prava koja najbolje opisuje njigovu zavisnost.

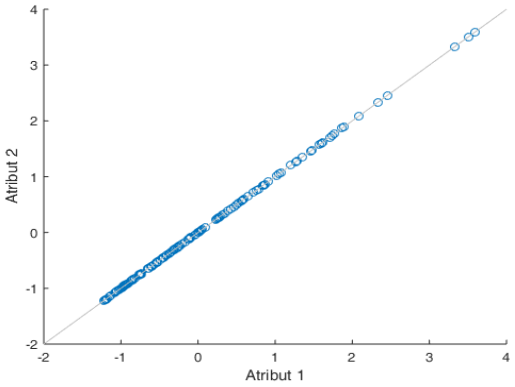
Na slici 2 je prikazan scatterolot atributa 11 i 14 prve klase pacijenata, što je maksimalna vrednost za klasu.



Sl. 2. Korelacija atributa 11 i 14 grupe obolelih.

Primećuje se da je korelacija ovih atributa gotovo 1, što se moglo očekivati s obzirom na njigovu jednaku prirodu – oba atributa su varijante shimmer–a. Stoga, jedan od ta dva atributa se može izbaciti iz razmatranja jer ne nosi informacije.

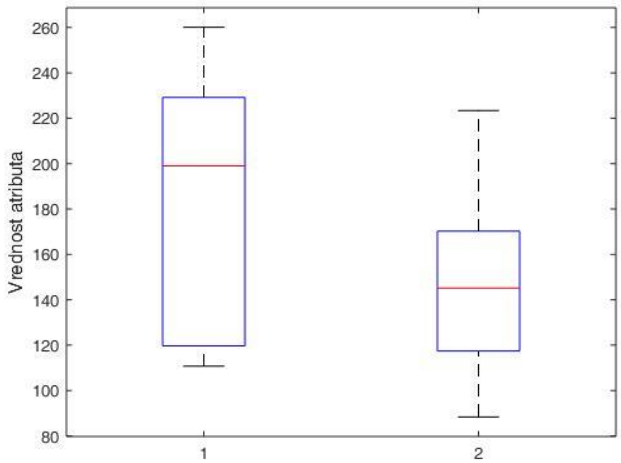
Kako i za obolele, isto se primećuje i za zdrave pacijente, što je prikazano na slici 3.



Sl. 3. Korelacija atributa 11 i 14 grupe zdravih.

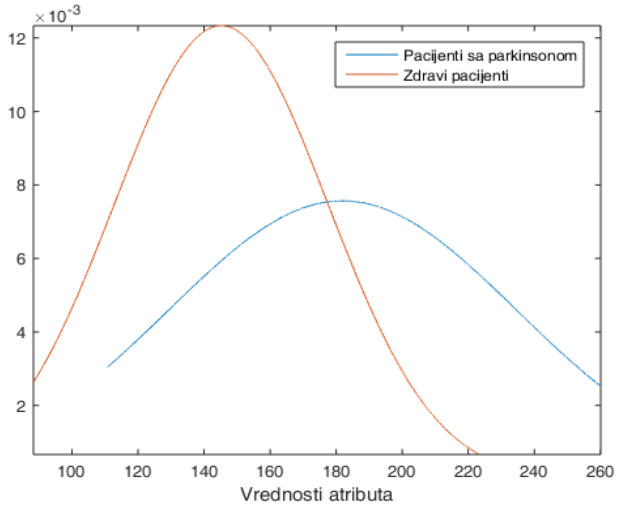
E. Poređenje boxplot-a i grafika normalne raspodele različitih grupa pacijenata

Urađene su sledeće analize prvog atributa (osnovne frekvencije glasa) svake od klase i direktno su upoređene dobijene vrednosti. Iscrtani su boxplot–ovi, slika 4, i grafici normalne raspodele, slika 5.



Sl. 4. Boxplot – ovi prvog atributa obe klase.

Analizom boxplot–ova mogu se uočiti razlike atributa u različitim klasama. Prvo, median za klasu I je značajno veći nego za klasu II. Drugo, interkvartilni opseg za klasu I je veći dok su dinamički opsezi približno isti. Takođe, i minimalna i maksimalna vrednost se primetno razlikuju dok outlier–i ne postoje ni za jednu klasu.



Sl. 5. Normalne raspodele prvog atributa za obe klase.

Sa grafika normalne raspodele se može uočiti nekoliko razlika atributa između klasa. Prva je razlika srednjih vrednosti, minimuma i maksimuma koji su veći kod obolelih pacijenata. Zatim, standardna devijacija obolelih pacijenata je veća tj. raspodela je više raširena.

Boxplot i grafika normalne raspodele daju sličan uvid u karakteristike atributa, pa su i zaključci slični. Razlika između klasa postoji i moglo bi se zaključiti da je korisna za klasifikaciju.

Međutim, problem se može javiti kao posledica prirode atributa. Osnovna frekvencija glasa je karakteristična za svaku osobu i zavisi od pola i godina. Parkinsonova bolest izaziva povećanje osnovne frekvencije, ali ukoliko ne znamo koliko je ona bila pre nastanka bolesti, ne možemo je uporediti sa trenutnom i sa sigurnošću znati razlog njene vrednosti. Npr. muškarci u proseku imaju nižu osnovnu frekvenciju od žena, nezavisno od toga da li boluju od Parkinsonove bolesti. Ukoliko se u obzir ne uzmu pol i starost pacijenata, rezultati analize ove karakteristike mogu biti nepouzdani.

U datoj bazi podataka se ipak može opravdati postojanje razlike kao posledice bolesti. U klasi obolelih pacijenata ima procentualno više muškaraca i prosečne godine su veće u odnosu na klasu zdravih. Ovakva struktura pacijenata bi trebala da snizi prosečne i ekstremne vrednosti osnovne frekvencije ali je slučaj obrnut. Stoga se zaključuje da je razlika atributa pouzdana i potencijalno korisna u klasifikaciji.

III. Zaključak

Za sada nisu dobijeni značajni rezultati jer nisu upoređena sva obeležja različitih grupa pacijenata.