PRIMENJENA PSIHOLOGIJA, 2011/4, STR. 335-351

UDC 159.9.072.5
Originalni naučni rad

Bojan Janičić[1] i Zdenka Novović

Odsek za psihologiju,
Filozofski fakultet,
Univerzitet u Novom Sadu

PROCENA USPEŠNOSTI U KLASIFIKOVANJU REZULTATA NA OSNOVU GRANIČNIH (CUT-OFF) SKOROVA: RECEIVER OPERATING CHARACTERISTIC CURVE[1]

Rezime

Cilj rada je da se ukaže na mogućnosti upotrebe ROC krive (engl. receiver operating characteristic curve) za utvrđivanje klasifikatornih mogućnosti testa. Objašnjeni su pojmovi senzitivnosti i specifičnosti koje leže u osnovi izrade ROC krive, a data su tumačenja i formule i za izračunavanje pozitivne i negativne prediktivne vrednosti, kao i tačnosti testa. ROC kriva je grafički prikaz senzitivnosti i specifičnosti za svaki mogući granični skor (rezultat na testu) u koordinatnom sistemu gde su na ordinati prikazane vrednosti senzitivnosti, a na apscisi vrednosti specifičnosti oduzete od 1. Objašnjeno je kako se na osnovu krive iz tabele svih vrednosti senzitivnosti i specifičnosti može odrediti optimalan granični skor za neki test ili za potrebe klasifikovanja druge vrste. Pokazano je kako se u statističkom programu SPSS unose podaci i analiziraju dobijeni rezultati ROC analize. Takođe su ponuđeni i drugi programi i paketi koji omogućavaju ovu analizu sa brojnim dodatnim mogućnostima. Na kraju je ukazano na rezultate istraživanja u okviru kliničke psihologije koji su utemeljeni na ROC analizi i karakteristikama testa, odnosno klasifikacije na kojima je ova analiza utemeljena.

Ključne reči: senzitivnost, specifičnost, ROC analiza, površina pod krivom (AUC)

[1] Adresa za korespodenciju:
janicic@ff.uns.ac.rs

Primljeno: 23.12.2011.
Prihvaćeno za štampu: 17.01.2012.

[1] Rad je nastao u okviru projekta „Nasledni, sredinski i psihološki činioci mentalnog zdravlja“ (broj 179006) koji finansira Ministarstvo prosvete i nauke RS

Uvod

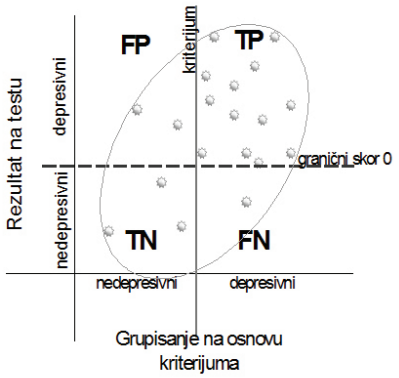
U istraživanjima, a još češće u praksi, postavljamo neka bitna pitanja na osnovu čijeg odgovora mogu zavisiti vrlo važne praktične odluke. Pitanja kojima se bavimo u ovom članku odnose se na uspešnost nekog testa u detekciji osoba sa određenim kvalitetom. Da li je instrument koji posedujemo uspešan u razdvajanju depresivnih i nedepresivnih, paranoidnih i neparanoidnih, psihopata od nepsihopata i slično? Odgovori na ova pitanja, koji nam prvi padaju na pamet, su svakako u vezi sa validnošću testa, npr. izračunavanjem korelacija između rezultata testa i kriterijuma. Međutim, mnogi naši instrumenti, koji se ne pokazuju validnim na ovaj način, mogu imati neke karakteristike zbog kojih opstaju u praksi. Na primer, Biro, Ristić, i Novović (1987) su pokazali da Roršahov metod u identifikaciji psihotičnih pacijenata, uprkos niske validnosti izračunate na osnovu korelacija, sadrži neke indikatore koji se javljaju samo kod psihotičnih pacijenata, te se, kada se oni jave, sa velikom sigurnošću može dijagnostikovati psihoza. Njihova validnost je niska jer se ovi indikatori retko pojavljuju. Otuda i njihovo odsustvo nema veliku prediktivnu moć, ali zato prisustvo ima.

Srodna pitanja, na koja nam ponekad ne odgovaraju standardni statistički postupci, odnose se na određivanje graničnih ili cut-off skorova. Do kog skora možemo tvrditi da osoba poseduje neku osobinu u niskom stepenu, a od kog u naglašenom? Kada je u pitanju dimenzionalna pojava, može nam biti potrebno i da odredimo više ovakvih graničnih rezultata, koji će određivati kategorije, npr. niske, umerene ili visoke depresivnosti. Iako se na osnovu deskriptivnih pokazatelja može odrediti koji su rezultati najčešći, a koji retki, na osnovu kog rezultata se populacija može podeliti na pola u odnosu na neki kvalitet, koji rezultat je prosečan za datu populaciju i slično, često nas interesuje i koji skor uspešno deli populaciju na one koji imaju i one koji nemaju mereni kvalitet. Mogućnosti adekvatne klasifikacije osoba sa različitim kvalitetima u odgovarajuće kategorije je, takođe, jedno od pitanja koje se može proveravati na osnovu statističkih postupaka o kojima će u ovom članku biti reči. Na primer, da li neki test depresivnosti uspešno razlikuje depresivne i anksiozne pacijente (videti primer ovakve analize kod nas u Novović i Janičić, 2009, ili u stranoj literaturi u Somoza, Steer, Beck, & Clark, 1994). Često nas, takođe, interesuje i da li granični skor predložen za jednu populaciju važi i za drugu?

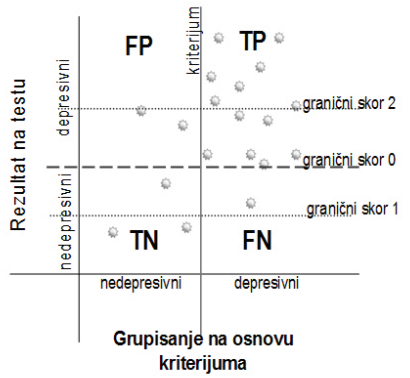
Senzitivnost, specifičnost i druge klasifikatorne karakteristike testa

Postoje statističke mere koje daju odgovore na ova pitanja. Pretpostavimo da imamo grupu depresivnih pacijenata koji su na osnovu baterije kliničkih testova di-

jagnostikovani kao depresivni poremećaj. Takođe, imamo i grupu ispitanika koji nisu depresivni, a ista baterija je to i potvrdila. Obe grupe rešavale su novi test depresivnosti sa datim graničnim skorom koji sugeriše da rezultati iznad njega govore u prilog ozbiljne, kliničke depresije. Rezultati obe grupe u odnosu na test, kao i kriterijum (bateriju kliničkih testova), prikazani su na skater dijagramu (Slika 1).



Slika 1. *Skater dijagram hipotetičkih rezultata na testu depresivnosti*



Slika 2. *Skater dijagram hipotetičkih rezultata sa različitim graničnim skorovima*

Rezultati ispitanika su na skateru prikazani kružićima. Za ispitanike levo od pune linije je na osnovu psihološke baterije testova (kriterijuma) ustanovljeno da nisu depresivni, a za one desno da jesu. Ispitanici iznad isprekidane linije su na osnovu novog testa depresivnosti označeni kao depresivni – skor im je iznad graničnog, a oni ispod nje su prema testu nedepresivni. U oblastima TP i TN nalaze se rezultati ispitanika kod kojih postoji slaganje između testa i kriterijuma. U prvom slučaju (TP), i na osnovu testa i na osnovu kriterijuma, oni su depresivni – ove rezultate zovemo zato *tačni pozitivni*. U drugom slučaju, i test i kriterijum ukazuju na to da ispitanici nisu depresivni – to su *tačni negativni* rezultati. U oblasti FP su ispitanici koje je test lažno "prepoznao" kao depresivne, a da oni to zapravo nisu (*falš pozitivni*), a u FN su depresivni ispitanici koje test nije prepoznao kao takve (*falš negativni*). Test koji ima dobro postavljen granični skor, treba da ima što manje falš negativnih rezultata – da bude što manje neprepoznatih depresivnih ispitanika, a da pri tome ni broj falš pozitivnih ne postane prevelik – da se na osnovu testa ne proglase depresivnim i oni koji to nisu. Koliko će test uspeti u ovom zadatku selekcije ne zavisi, naravno, samo od graničnog skora, već, pre svega, od validnosti samog testa. Što je test kriterijumski validniji za osobinu koja se meri, biće manje falš negativnih i falš pozitivnih rezultata. Zamišljena elipsa koja obuhvata rezultate ispitanika na skater dijagramu (Slika 1) biće uža, a samim tim će biti manje rezultata koji će se nalaziti u oblastima FN i FP. Takođe, uspešnost

klasifikacije na osnovu nekog testa zavisiće i od bazne stope (prevalence pojave u populaciji) i toga koliko je kriterijum koji smo koristili bio oštar (koliko je baterija precizna) u odvajanju stvarno depresivnih, kao i od prirode same pojave i prirode njene merljivosti (koliko se precizno može izmeriti, koliko se ljudi razlikuju po datoj osobini). Međutim, kada je merenje već obavljeno, to su problemi na koje ne možemo uticati, ali možemo da menjamo granični skor na testu. Šta se dešava ako ovaj skor snizimo?

Na Slici 2 ova situacija je prikazana graničnim skorom 1. U ovoj situaciji, ispod graničnog skora ostaje manji broj rezultata, a povećava se broj rezultata iznad graničnog skora. Snižavanjem graničnog skora smanjuju se šanse da test ne prepozna osobe koje su zaista depresivne (povećaće se broj tačnih pozitivnih i smanjiti broj falš negativnih). Snižavanjem kriterijuma povećali bismo *osetljivost tj. senzitivnost testa* (Glaros & Kline, 1988).

Senzitivnost je definisana kao sposobnost instrumenta ili baterije da pokaže pozitivni rezultat kod osoba koje zaista poseduju dijagnostikovani kvalitet koji nas interesuje. Ona se izražava kao proporcija osoba sa datom osobinom, koje su na instrumentu postigle rezultat iznad graničnog skora, tj. kao proporcija tačnih pozitivnih rezultata u odnosu na sve zaista depresivne osobe (Glaros & Kline, 1988). Ako bismo ovo pokušali da izrazimo preko oblasti na našem skater dijagramu koje smo obeležili sa TP, FP, TN i FN, senzitivnost bismo izrazili formulom TP / (FN + TP). Ako prebrojimo rezultate koji pripadaju ovim oblastima, možemo izračunati, na našem primeru, kolika je senzitivnost testa sa graničnim skorom koji je postavljen kao na Slici 1. Tačnih pozitivnih je 13 (TP), a pored njih još 1 rezultat je desno od kriterijuma, tj. jedna osoba koja je zaista depresivna nije se takvom pokazala na testu (FN). Prema formuli, rezultat bi bio 13 / 14 = 0.92.

Videli smo kako se menjanjem graničnog skora menja broj rezultata u svim oblastima na graficima 1 i 2. Na grafiku 2. granični skor 2 je viši od predviđenog (0). Jasno je da time snižavamo senzitivnost testa, jer se smanjuje broj depresivnih subjekata koji su prepoznati kao takvi na testu. Time se povećavaju šanse da neko ko je zaista depresivan, testom ne bude identifikovan kao takav. Povišavanjem graničnog skora test lakše ”promašuje“ subjekte koji su depresivni, ali zato sa većom sigurnošću možemo da tvrdimo da su oni koji imaju rezultat iznad ovog skora zaista depresivni. Takođe, kad je granični skor visoko postavljen, možemo biti veoma sigurni da smo većinu osoba bez merenog kvaliteta na osnovu testa prepoznali kao takve. Drugim rečima test postaje *specifičniji.* Specifičnost se definiše kao sposobnost instrumenta da pokaže negativni rezultat kod osoba koje ne poseduju merenu osobinu. Izražava se kao proporcija osoba bez merenog kvaliteta koje postižu rezultate na instrumentu ispod cut-off skora, tj. negativan rezultat, ili kao proporcija tačnih negativnih rezultata u odnosu na sve one koji ne poseduju mereni kvalitet (Glaros & Kline, 1988). Našim simbolima možemo izreći i formulu za specifičnost: TN / (FP + TN). Tako saznajemo da naš test sa

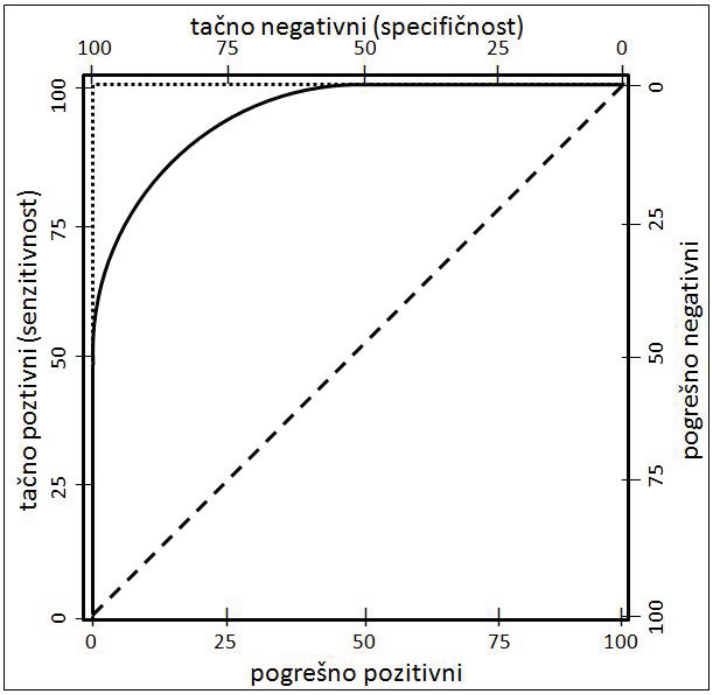
graničnim skorom 0 (Slika 1) ima specifičnost od 3 / 5 = 0.60. Iz primera sa povišavanjem i snižavanjem graničnog skora (Slika 2), jasno nam je da sa snižavanjem kriterijuma senzitivnost raste, a specifičnost opada, i obrnuto. U zavisnosti od namene testa i svrhe za koju se trenutno upotrebljava, možemo granični skor postaviti tako da nam što manje "promaknu" stvarno depresivni, ili da što više nedepresivnih ljudi bude takvo i na osnovu našeg testa.

Do sada smo videli kako na osnovu tabele 2 x 2 možemo jednostavno da izračunamo senzitivnost i specifičnost testa sa datim graničnim skorom. Iz iste tabele možemo saznati još neke karakteristike testa kao što su *preciznost - prediktivna vrednost pozitivnog rezultata* (TP / TP + FP), što je proporcija tačno dijagnostikovanih subjekata u odnosu na sve pozitivne rezultate. Ona govori o verovatnoći da pozitivni rezultat zaista meri osobinu koju treba da meri, odnosno govori o poklapanju testa i kriterijuma kada su u pitanju pozitivni rezultati. Ako je test 100% precizan, onda su svi subjekti sa pozitivnim rezultatom zaista depresivni. Takođe, možemo izračunati *prediktivnu vrednost negativnog rezultata* – TN / (TN + FN) koja ukazuje na to kolika je verovatnoća da je osoba koja je dobila negativan rezultat zaista bez osobine koju merimo. Prediktivna vrednost pozitivnog i negativnog rezultata zavise od broja osoba sa datim kvalitetom u populaciji. Ako je u pitanju broj osoba sa poremećajem, rekli bismo da se prediktivne vrednosti menjaju sa prevalencijom poremećaja (Akobeng, 2007). Konačno, *tačnost testa* ili u kojoj meri je naš test sa datim graničnim skorom generalno verodostojan, u smislu njegovog poklapanja sa prisustvom pojave u populaciji koje smo ustanovili, u našem slučaju, na osnovu baterije i prihvaćenog kriterijuma, izračunava se tako što sve tačne rezultate (TP i TN) podelimo sa ukupnim brojem rezultata (TP + TN + FP + FN).

ROC analiza[1]

Pretpostavimo sada da nemamo granični skor za neki test i da želimo da ga odredimo, tako da postignemo najbolji odnos senzitivnosti i specifičnosti, odnosno da proporcija TP i proporcija TN bude najveća moguća. Ovo nije jednostavno, jer kako je rečeno, povišavanje jedne mere povlači sniženje druge. Kada bismo želeli da ovakvu odluku donesemo na osnovu izračunavanja senzitivnosti i specifičnosti za svaki mogući granični skor, dobili bismo dug niz paralelnih vrednosti i njihovim upoređivanjem mogli bismo ustanoviti gde je senzitivnost dovoljno visoka, a da je pri tome specifičnost najmanje oštećena. Statistička tehnika koja za cilj ima utvrđivanje granične vrednosti nekog testa koji daje najbolji odnos specifičnosti i senzitivnosti naziva se analiza ROC krive.

[1] Engl. Receiver Operating Characteristic Curve (ROC kriva) je metoda grafičkog prikazivanja odnosa senzitivnosti i specifičnosti koja je prvi put primenjena tokom II svetskog rata za analizu radarskih signala, da bi u naučnu literaturu ušla preko Teorije detekcije signala (Krzanowski & Hand, 2009).

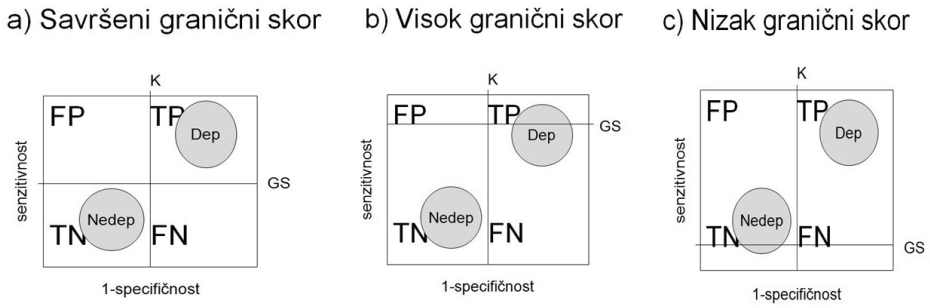


Slika 3. *Ose ROC krive*

ROC kriva je grafički prikaz senzitivnosti i specifičnosti za svaki mogući granični skor (rezultat na testu) u koordinatnom sistemu gde su na ordinati (y) prikazane vrednosti senzitivnosti (proporcija tačnih pozitivnih), a na apscisi (x) vrednosti specifičnosti oduzete od 1 (1 - specifičnost)[2], čime se dobija proporcija falš pozitivnih rezultata (Slika 3). Na Slici 3 vidimo u kakvom odnosu stoje proporcije četiri osnovne veličine iz ćelija 2 x 2 matrice (TP, FP, TN, FN).

Ako je pojava koju merimo takva da se oni koji poseduju kvalitet i oni koji ga ne poseduju uopšte ne razlikuju (zamislimo da se populacije depresivnih i nedepresivnih uopšte ne razlikuju po kvalitetu koji merimo), onda naš test ima šanse 50:50% da slučajno "pogodi" ko je depresivan, i koji god granični skor na testu da uzmemo, proporcija ispravno klasifikovanih ostaje ista. Ova situacija bi na ROC grafiku bila predstavljena dijagonalnom linijom koja spaja donji levi i gornji desni ugao, odnosno dve nulte tačke sa Slike 3. Ova dijagonala se obično zove *dijagonala slučajnog ishoda* (engl. chance diagonal) i prikazana je isprekidanom linijom na Slici 3 (Krzanowski & Hand, 2009).

[2] Na apscisi može biti predstavljena i specifičnost, ali su tada vrednosti na osi opadajuće i idu od 1 do 0.



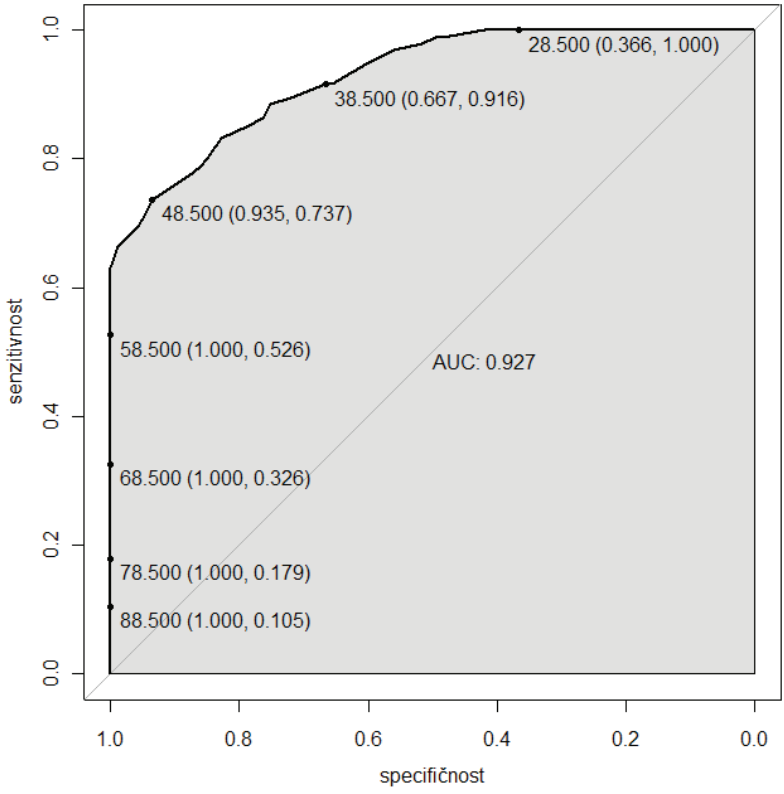
Slika 4. *Šta se dešava sa rezultatima kada se menja granični skor, a populacije onih sa i bez kvaliteta se savršeno razlikuju?*

Sa druge strane, zamislimo situaciju u kojoj bi svi depresivni pacijenti bili jasno različiti od nedepresivnih, a test je dobar indikator depresivnosti. Ako krenemo od najviših graničnih skorova na testu, senzitivnost – sposobnost da se prepoznaju depresivni pacijenti, će biti niska, ali će i falš pozitivni biti odsutni (grafik b na Slici 4) i većina rezultata će biti falš negativna. Na ROC krivoj taj rezultat će se nalaziti na samoj *y*-osi (jer su falš pozitivni 0) i to na nižim vrednostima za senzitivnost. Svaki sledeći niži granični skor na ROC grafiku davaće presek senzitivnosti i (-) specifičnosti na samoj *y*-osi, senzitivnost će se sa spuštanjem kriterijuma povećavati (test će sve više depresivnih ljudi prepoznati kao takve), pa tako sve do savršenog kriterijuma, kada kriva stiže do gornjeg levog ugla gde se sustiču tačke maksimalne senzitivnosti i specifičnosti. Sa daljim spuštanjem graničnog skora počinju da rastu falš pozitivni, a opadaju tačno negativni rezultati, odnosno specifičnost se smanjuje (grafik c na Slici 4) i ROC linija prati gornju liniju grafika. Dakle, idealni test kojim merimo jasno različite kvalitete dao bi ROC krivu koja prati levu i gornju ivicu grafika, kao što je prikazano tačkastom linijom na Slici 3.

ROC krive, u najvećem broju slučajeva, se nalaze negde između dve opisane situacije, između dijagonale slučajnog ishoda i krive koja se poklapa sa osama senzitivnosti i specifičnosti. Što je dobijena krivulja bliže idealnoj, tačkastoj sa Slike 3, to je test diskriminativniji – bolje razlikuje dve grupe, jer se i senzitivnost i specifičnost za svaki granični skor približavaju idealnim vrednostima i veće su šanse da se među njima pronađe onaj granični skor koji će imati mali broj i falš pozitivnih i falš negativnih. Suprotno, što je krivulja bliže dijagonali slučajnog ishoda, test je manje diskriminativan i njegove klasifikatorne mogućnosti se sve manje razlikuju od slučajnog pogađanja.

Iz navedenog sledi da, što je površina koju zahvata ROC kriva veća, to je u pitanju diskriminativniji test. Tako dolazimo do pojma *površine ispod ROC krive* (engl. Area Uder the Curve: AUC) koja je sastavni deo ROC analize. Interpretacija ove površine, koja se najčešće koristi, je da je AUC pokazatelj verovatnoće da će na osnovu testa viši skor za slučajno odabranog ispitanika imati osoba sa datim kva-

litetom, nego osoba bez kvaliteta (Krzanowski & Hand, 2009). Kao takva, AUC je ekvivalentna (WIlcoxon-) Mann-Whitney U testu (Sing, Sander, Beerenwinkel, & Lengauer, 2005). Prevedeno na primer sa depresivnošću, AUC govori o verovatnoći da će slučajno odabrana depresivna osoba na testu depresivnosti imati viši skor, nego osoba koja nije depresivna.[3] Na Slici 5 je prikazana ROC kriva jednog psihološkog testa. Označene su neke od mogućih graničnih vrednosti, a u zagradi pored njih su date specifičnost i senzitivnost. Kriva kreće iz donjeg levog ugla gde su granični skorovi viši, senzitivnost niska, a specifičnost visoka. Kako idemo ka gornjem levom uglu, granični skorovi su sve niži, senzitivnost sve viša, a specifičnost sve niža. Kao optimalan granični skor označen je 48.5. AUC je zatamnjena i iznosi 0.927, što nam govori da je verovatnoća 92.7% da će depresivna osoba imati viši skor na testu od one koja to nije.

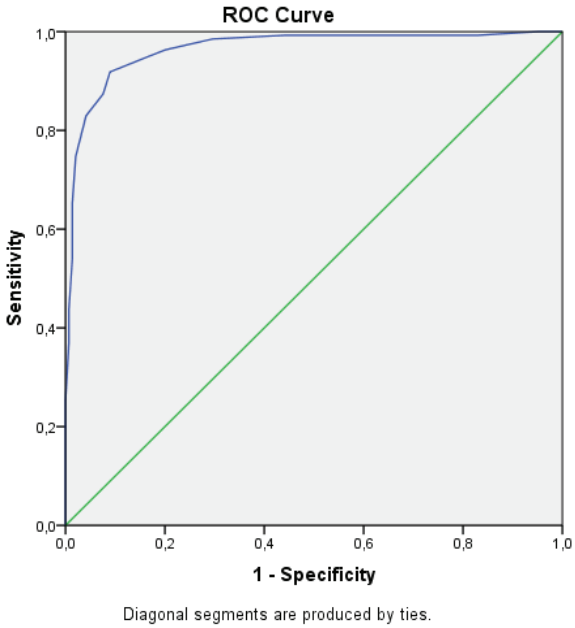


Slika 5. *ROC kriva sa označenim odabranim graničnim vrednostima*

[3] Na osnovu definicije AUC, teorije verovatnoće i računanja, te činjenice da je najveća moguća AUC = 1, sledi da je AUC prosečna proporcija tačnih pozitivnih rezultata koji se dobijaju ravnomerno iznad svih mogućih proporcija falš negativnih rezultata u rasponu od 0 do 1 (Krzanowski & Hand, 2009).

Kako u okviru statističkih paketa analizirati ROC krivu?

Analiza ROC krive se teško može zamisliti bez računara. Jedan od poznatijih statističkih paketa koji nudi mogućnost analize ROC krive je IBM SPSS Statistics. Za analizu ROC krive u ovom, ili bilo kom drugom statističkom paketu, potrebno je imati dve grupe subjekata koje su grupisane prema nekom kriterijumu, ili kako se to često naziva ”zlatnom standardu“. Prikazaćemo korake analize na primeru analize skraćene D skale MMPI-a, kada smo prikupili rezultate ove skale od 141 pacijenta, prosečne starosti od 47.35 godine, koji su se lečili od nekog od kliničkih oblika depresije na klinici u Novom Sadu (detalji odabiranja stavki za skalu mogu se videti u Novović i Biro, 2009). Kontrolna grupa od 144 nedepresivnih subjekata, prosečne starosti 43.53 godine, prikupljena je iz opšte populacije i takođe je popunjavala isti test. Kriterijum zlatni standard je bio psihijatrijska dijagnoza. Rezultati subjekata obe grupe su uneti u matricu. Pored varijable sa skorovima na testu (nazovimo je ”D skorovi“), formirana je i varijabla ”kriterijum“ u kojoj je svim ispitanicima dodeljen broj 1 ako su u pitanju pacijenti, a 0 ako su iz opšte populacije. U SPSS-u željenu analizu nalazimo u meniju ”Analyze“, gde biramo opciju ”ROC Curve“. U dijalogu koji se tada pojavi, potrebno je definisati test varijablu (D skorovi) i varijablu stanja (kriterijum). Zatim je u polju ”Value of state variable“ neophodno izabrati i vrednost varijable stanja koja označava pozitivne slučajeve – u našem slučaju brojem 1 su označene osobe koje su depresivne prema kriterijumu, pa smo upisali br. 1. Takođe, možemo odabrati detalje analize koje želimo da nam program prikaže (grafik sa ili bez dijagonale slučajnih ishoda, standardnu grešku i interval poverenja za AUC, te koordinate krive). Na Slici 6 vidimo ROC krivu koju smo dobili u ispisu rezultata. Već na prvi pogled vidimo da se dobijena kriva približava levoj i gornjoj liniji dijagrama, te već na osnovu nje vidimo da je naša skraćena D skala dobro klasifikatorno sredstvo u razdvajanju depresivnih pacijenata i opšte populacije.



Slika 6. *ROC kriva dobijena na osnovu rezultata ispitivanja klasifikacije pomoću skraćene D skale MMPI-a.*

U Tabeli 1 nalaze se rezultati u vezi AUC za naš primer. Na osnovu veličine površine (Area = 0.96) saznajemo da će depresivni pacijent imati 96% veće šanse nego osoba iz opšte populacije da ima povišen skor na D skali.

Tabela 1.
SPSS ispis rezultata u vezi AUC na primeru skraćene D skale MMPI

Area Under the Curve

Test Result Variable(s): skracena D skala

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.965	.011	.000	.944	.986

Ostale vrednosti iz tabele su standardna greška (0.011), pokazatelj značajnosti rezultata (p = 0.000) i interval pouzdanosti (0.94 - 0.99).

Pored krive i tabele sa AUC parametrima, SPSS daje i tabelu sa koordinatama krive u kojoj postoje tri kolone. U prvoj (”Positive if Greater Than or Equal To“) su svi mogući rezultati testa kao potencijalni granični skorovi. U našem slučaju, minimum je 0, a maksimum 20, te postoje 20 ovih rezultata. U drugoj koloni date su senzitivnosti za svaki od rezultata, a u trećoj 1 - specifičnosti, odnosno proporcija falš pozitivnih rezultata, opet za svaki rezultat testa, tj. granični skor. Što je potencijalni granični skor iz prve kolone niži, više su i senzitivnost i 1 - specifičnost. Potrebno je da nađemo onaj granični skor koji podrazumeva da su i senzitivnost i specifičnost najviši (i proporcija tačnih pozitivnih i tačnih negativnih je najviša moguća). Optimalan granični skor ćemo naći kad nađemo skor za koji se dobija najviša razlika kada se od senzitivnosti oduzme 1 - specifičnost. U našem primeru pronašli smo da se kod skora 12 nalaze vrednosti sa najvećom razlikom, te smo zaključili da je ovo optimalan granični skor. Senzitivnost za ovaj skor je iznosila 0.919, a 1 -specifičnost 0.085, što kad oduzmemo od 1, daje specifičnost od 0.915. Zaključili smo da sa graničnim skorom 12, naš test daje 91.9% tačnih pozitivnih rezultata i 91.5% tačnih negativnih, pa smo ga predložili kao optimalan za razlikovanje klinički depresivnih osoba, od nedepresivnih iz opšte populacije.

Identičan postupak i izlaz analize imamo i u besplatnom statističkom paketu PSPP[4] koji predstavlja ”open-source“ alternativu paketu SPSS. Ni jedan, ni drugi ne predlažu optimalnu cut-off vrednost.

Naravno, optimalan granični skor možemo i ovde sami odrediti tako što ćemo pronaći graničnu vrednost u kojoj je suma senzitivnosti i specifičnosti najveća (u tabeli sa koordinatama ROC krive). Grafički, optimalna granična vrednost nalazi se na mestu gde je kriva najudaljenija od dijagonale slučajnih ishoda i predstavlja tangentu na ROC krivu pod uglom od 45 stepeni.

Analizu ROC krive možemo obraditi i u R-u, koji više predstavlja (besplatno) softversko okruženje za statistička izračunavanja, nego statistički paket poput SPSS-a ili Statistica-e. Za razliku od konvencionalnih statističkih paketa, R nema bogat grafički korisnički interfejs, pa korisnik većinu komandi unosi na komandnoj liniji. Iako je to nešto na šta možda nismo navikli, to je upravo ono što nam omogućava da analize prilagodimo svojim potrebama.

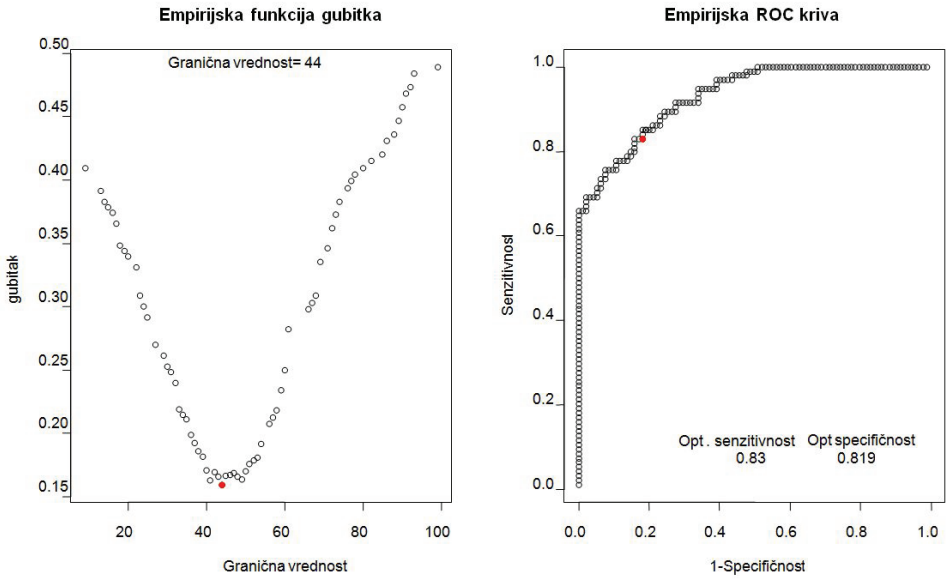
U R-u postoji veći broj paketa namenjenih ROC analizi: ThresholdROC, pROC, ROCR, ROCwoGS, rocplus (Wheeler, B. 2011, Wheeler, R.E. 2011)…[5] Većina ovih paketa ima mogućnost pronalaženja optimalne granične vrednosti. S obzirom na ovu mogućnost treba izdvojiti paket ThresholdROC[6] (Skaltsa, 2011) koji izračunava optimalne granične vrednosti kada imamo 2 ili 3 grupe subjekata (2

[4] http://www.gnu.org/software/pspp/get.html

[5] Po instalaciji R-a potrebno je željeni paket instalirati, tj. «downloadovati» preko menija „Packages“ i opcije „install packages“. Pre nego što krenete u prvu analizu, potrebno je željeni paket učitati komandom „library (NazivPaketa)“, vodeći računa o velikim i malim slovima koje R razlikuje. Nakon toga, pomoć vam je dostupna ako ukucate komandu „?NazivPaketa“.

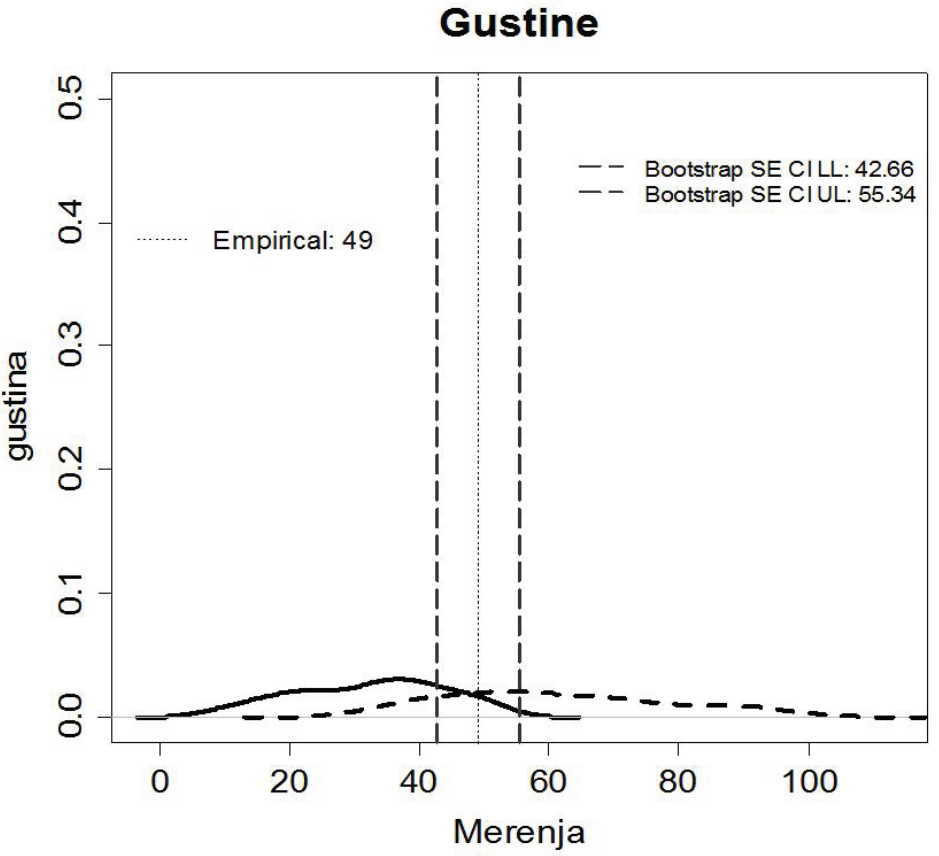
[6] http://cran.r-project.org/web/packages/ThresholdROC/index.html

ili 3 stanja), bez obzira imaju li grupe jednake ili različite varijanse i bez obzira na distribuciju. Takođe, ovaj paket može da izračuna optimalnu vrednost graničnog skora na osnovu minimizacije *funkcije gubitka* (engl. cost function), koja uzima u obzir senzitivnost, specifičnost, prevalencu i težine (pondere) koje po potrebi možemo dodeliti proporcijama TP, TN, FP i FN (Slika 7). Gubitak će biti veći što su specifičnost i senzitivnost manje, a pojedinačne proporcije će doprinositi njenom rastu, što im veći ponder dodelimo. Ograničenje je da pogrešne klasifikacije (FP i FN) ne smeju imati manje pondere od ispravnih (TP i TN), što je i u skladu sa zdravim razumom (Skaltsa, Jover, & Carrasco, 2010).



Slika 7. *Grafikon funkcije gubitka i ROC kriva iz R paketa ThresholdROC*

U ovom paketu moguće je izračunati i *standardne greške procenjene granične vrednosti* upotrebom bootstrapping procedure i na osnovu njih formirati interval poverenja (Slika 8).



Slika 8. *Razdvajanje grupa na osnovu empirijske granične vrednosti i interval poverenja granične vrednosti (ThresholdROC)*

Od ostalih paketa, po raznolikosti grafikona i pokazatelja koje na njima može da prikaže, izdvaja se paket ROCR (Sing, Sander, Beerenwinkel, & Lengauer, 2005). Paket ROCwoGS (Wang, Turnbull, Grohn, & Nielsen, 2007) namenjen je utvrđivanju optimalnog graničnog skora kada nemamo dobar kriterijum (zlatni standard), a pROC (Robin et al., 2011) omogućava poređenje dve ROC krive.

Primena u kliničkoj psihologiji

Provera uspešnosti klasifikovanja i određivanje graničnih vrednosti je u medicinskim naukama ustaljena i skoro neizostavna metoda. ROC analizom se određuje koliko se uspešno pacijenti na osnovu pozitivnog rezultata na nekom testu razlikuju od zdravih, da li se na osnovu nekog dijagnostičkog sredstva mogu uspešno razdvajati neke vrste ili podvrste poremećaja, ili se određuje nivo nekog terapijskog sredstva kojim se mogu postići maksimalni efekti sa minimumom štete. U psihologiji, osim kod zadataka diskriminacije u eksperimentalnoj oblasti odakle je metoda i potekla, druge mogućnosti primene analize klasifikacije preko ROC krive su relativno kasno uočene. Tek u poslednje vreme raste primena ove analize u kliničkoj psihologiji. Pintea & Moldovan (2009) u uvodnom delu članka koji se bavi mogućnostima primene ROC analize upravo u kliničkoj psihologiji, navode izvestan broj provera korisnosti i dijagnostičkih moći skrining instrumenata uz pomoć ove analize. Tako smo saznali da su u poslednjih pet godina ROC analizom proverene karkteristike testova za identifikaciju budućih teškoća u shvatanju čitanja, poremećaja baziranih na zloupotrebi alkohola i droga, neuropsihološkog oštećenja, depresije, opsesivno-kompulzivnog poremećaja, bipolarnog poremećaja, suicidalnog rizika, demencije, rizika za odustajanje od različitih tretmana itd. Srpski psiholozi mogli su se do sada upoznati sa mogućnostima ROC metode, kao i statističkih mera koje im leže u osnovi, u određivanju graničnih vrednosti novih testova (pogledati poglavlja koja se odnose na merenje psihopatoloških fenomena u Biro, Smederevac, i Novović, 2009), u utvrđivanju diskriminativnih mogućnosti testova u razlikovanju dve vrste poremećaja (Novović i Janičić, 2005), kao i u opisivanju i potkrepljivanju korisnosti jednih ili beskorisnosti drugih testova (Biro i sar, 1987; Novović, Ćulibrk, Dubovska, i Mišić-Pavkov, 1993). Ovaj članak imao je funkciju da olakša upotrebu i dâ novi podstrek za korišćenje ovih statističkih mera.

Literatura

Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Pædiatrica, 96*, 338–341.

Biro, M., Ristić, J. i Novović, Z. (1987). Procena verovatnoće predikcije kao pokušaj adekvatnije validacije Rorschacha. *Primjenjena psihologija*, *8*, 265-273.

Biro, M., Smederevac, S. i Novović, Z. (2009). *Procena psiholoških i psihopatoloških fenomena*. Beograd: Centar za primenjenu psihologiju.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44,* 1013–1023.

Krzanowski W. J., & Hand D. J. (2009). *ROC curves for continuous data*. Boca Raton, FL: CRC/Chapman and Hall.

Novović, Z., i Biro, M. (2009). Procena simptoma depresivnosti, U M. Biro, S. Smederevac, i Z.

Novović (Ur.), *Procena psiholoških i psihopatoloških fenomena*. Beograd: Centar za primenjenu psihologiju.

Novović, Z., Ćulibrk, L., Dubovska, M. i Mišić-Pavkov, G. (1993). Validacija Lišerovog kolor testa na uzorku depresivnih bolesnika. *Psihologija*, *26*, 167-172.

Novović, Z. i Janičić, B. (2005). Diskriminativne mogućnosti Hamiltonove skale depresivnosti: ROC analiza. *Psihologija*, *38*, 473-489.

Novović, Z. i Janičić, B. (2009). Objektivna procena depresivne kliničke slike. U M. Biro, S. Smederevac, i Z. Novović (Ur.), *Procena psiholoških i psihopatoloških fenomena*. Beograd: Centar za primenjenu psihologiju.

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9*, 49-66.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez J-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. doi:10.1186/1471-2105-12-77 Retrieved from http://www.biomedcentral.com/1471-2105/12/77, 20.12.2011.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics, 21,* 3940-3941.

Skaltsa, K. (2011). *ThresholdROC: Optimum threshold estimation based on cost function in a two and three state setting. R package version 1.0*. Retrieved from http://CRAN.R-project.org/package=ThresholdROC, 20.12.2011.

Skaltsa, K., Jover, L., & Carrasco, J. L. (2010). Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. *Biometrical Journal*, *52*, 676–697.

Somoza, E., Steer, R. A., Beck, A. T., & Clark, D. A. (1994). Differentiating major depression and panic disorders by self-report and clinical rating scales: Roc analysis and information theory. *Behaviour Research and Therapy*, *32,* 771-782.

Wang, C., Turnbull, B. W., Grohn, Y. T., & Nielsen, S. S. (2007). Nonparametric estimation of ROC curves based on bayesian models when the true di-

sease state is unknown. *Journal of Agricultural, Biological and Environmental Statistics,12,*128-146.

Wheeler, B. (2011). *ROC, Precision-Recall, Convex Hull and other plots*. Retrieved from http://cran.r-project.org/web/packages/rocplus/rocplus.pdf, 20.12.2011.

Wheeler, R.E. (2011). *Rocplus. The R project for statistical computing*. Retrieved from http://www.r-project.org/ 20.12.2011.

Bojan Janičić & Zdenka Novović

Department of Psychology, Faculty of Philosophy, University of Novi Sad

EVALUATION OF THE SUCCESS OF CLASSIFICATION BASED ON CUT-OFF SCORES: RECEIVER OPERATING CHARACTERISTIC CURVE

Abstract

Aim of this study to draw attention to possibilities for use ROC curve analysis (receiver operating characteristic curve) for determining the classification capabilities of the tests. Concepts of sensitivity and specificity, underlying creation of ROC curves, are explained. Interpretation of formulas for calculating the positive and negative predictive values and accuracy of the tests are also given. ROC curve is a graphical representation of sensitivity and specificity for every possible threshold score (test result) in the coordinate system where the ordinate shows the values of sensitivity and the abscissa value of 1-specificity. It is explained how to determine optimal threshold score on the basis of sensitivity and specificity, and how to perform ROC analysis in several statistical packages (SPS, PSPP and R). In the end, it is pointed to the findings within clinical psychology that are based on ROC analysis and test characteristics (such as sensitivity and specificity) on which this analysis is based.

Keywords: sensitivity, specificity, ROC analysis, area under curve (AUC)