Pristupi strojnog učenje u uvjetima neuravnoteženih podataka

Kovačić, Maja

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Maja Kovačić

PRISTUPI STROJNOG UČENJA U UVJETIMA NEURAVNOTEŽENIH PODATAKA

Diplomski rad

Voditelj rada:
dr. sc. Tomislav Šmuc

Zagreb, 2019.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Ovaj rad posvećujem svojim roditeljima i sestri jer su uvijek vjerovali u mene na ovome putu. Posebna zahvala ide momu Anti na svakodnevnoj podršci i neiscrpnom poticanju na napredak, te mojoj cimerici Maji bez koje bi ovo studiranje bilo samo studiranje. Najveća zahvala ide momu mentoru, dr. sc. Tomislavu Šmucu bez čijeg znanja i motivacije ovaj rad ne bi bio moguć.

Sadržaj

Uvod

Postoji više aspekata koji bi mogli utjecati na rezultate postojećih sustava učenja. Jedan od tih pogleda, povezan je s učenjem u uvjetima neuravnoteženih podataka, tj. uvjetima kada broj podataka koji pripadaju jednoj klasi daleko nadmašuje broj podataka u drugim klasama. Ovaj problem je poznat kao problem neuravnoteženih klasa i često se navodi kao prepreka uvođenja klasifikatora algoritama strojnog učenja. U takvoj situaciji, koja je rijetka, a stoga i bitna u stvarnom svijetu, sustav za učenje bi mogao imati poteškoće prilikom učenja koncepta pripadnosti klasi koja je u manjini. Takva vrsta podataka se nalazi, na primjer, u medicinskim bazama podataka o rijetkim bolestima, podacima za detektiranje anomalija, prepoznavanje lica i mnogim drugima. Rezultati eksperimenta, temeljeni na diskriminaciji induktivne sheme, sugeriraju da problem nije isključivo uzorkovan neravnotežom klasa već je također povezan s brojem podataka koji se preklapaju među klasama. Međutim, preklapanje klasa se obično javlja kod skupova podataka s asimetričnom i iskrivljenom distribucijom. Evaluacijski kriterij može odbaciti primjer manjinske klase kao šum i inducirani klasifikator može izgubiti svojstvo klasifikacije. U rješavanju ovog problema postoji više različitih pristupa [14]. Prvo, uvesti mjeru značajnosti pozitivnih primjera u postojeće algoritme ili kreirati nove uzimajući u obzir navedenu mjeru. Drugo, balansirati podatke prije samog procesa smanjujući efekt iskrivljenosti distribucije klasa. I posljednje, korištenje troškovno-osjetljivih metoda koje kombiniraju razinu algoritma i razinu podataka uključujući trošak pogrešne klasifikacije. Takvi rezultati su potaknuli dva nova načina rješavanja problema učenja u uvjetima neuravnoteženih klasa. Ovdje promatramo nekoliko metoda koje se bave ovim uvjetima, povezujući metodu prekomjernog uzorkovanja s metodama čišćenja podataka. Glavna motivacija metoda nije samo uravnoteženje podataka, nego i uklanjanje šumnih primjera koji leže na pogrešnoj strani granice odluke. Uklanjanje šumnih primjera može pomoći u boljem definiranju klasa, stoga omogućuje stvaranje jednostavnijih modela s boljim mogućnostima generalizacije. Zaključeno je da metode prekomjernog uzorkovanja mogu pomoći u indukciji klasifikatora koji su precizniji od onih koji su uzorkovani metodom pod-uzorkovanja. Naše dvije predložene metode dobro funkcioniraju u praksi, posebno za skupove podataka s malim brojem pozitivnih primjera. Slijedeći pristup za rješavanje ovog pitanja je korištenje ansambla. Ansambli klasifikatora u strojnom učenju služe za povećanje točnosti jednog klasifikatora trenirajući više njih te

kombinirajući njihove rezultate. U literaturi se obično termin „ansambl“ veže uz kolekciju klasifikatora koji je manja varijanta jednog, istog klasifikatora. Kada se formiraju ansambli, važna je njihova dosljednost sa skupom za treniranje te raznolikost klasifikatora. To je ono što ih čini točnima. U klasifikaciji, različitost još uvijek nije dovoljno razrađena, no ona je nužna, postoji nekoliko načina da se ona postigne. U ovom radu ćemo koristiti ansamble bazirane na učenju nad različitim distribucijama, kao što su Bagging i Boosting. Niti jedna od navedenih tehnika učenja ne rješava problem nebalansiranosti podataka sama. Štoviše, kombinacija s ostalim tehnikama dovodi do nekoliko prijedloga s pozitivnim rezultatima. Prednost takvih hibridnih pristupa jest što ne trebaju mijenjati bazni klasifikator. Modifikacija ansambla obično uključuje pred-obradu podataka prije učenja svakog klasifikatora. Mnogi radovi imaju razvijene studije o prikladnosti tehnika pred-obrade podataka. Ovdje su razmatrane različite familije algoritama ovisno o ansamblima i tehnikama za balansiranje podataka, te iznesena empirijska usporedba performansi ansambla s ciljem pronalaska klasifikatora koji daje najbolje ponašanje u odnosu na ostale i promatranja prikladnosti povećanja kompleksnosti klasifikatora korištenjem ansambla. Posljednjih godina, zajednica koja se bavi strojnim učenjem, sve je svjesnija potrebe za statističkom validacijom dobivenih rezultata. To se može pripisati zrelosti područja, sve većem broju stvarnih primjena i dostupnosti software-skih platformi koje olakšavaju razvoj novih algoritama ili poboljšanje starih, te njihovo međusobno uspoređivanje. U tipičnom članku strojnog učenja, novi algoritam ili neki njegov dio ili novi korak prije ili poslije procesa implicitno znači da takva promjena dovodi do poboljšanja performansi postojećeg algoritma. Alternativno, različita rješenja problema su predložena i potrebno je razdvojiti uspješne od neuspješnih. Naredno poglavlje istražuje poveznicu između teorijskog i praktičnog rada. Usporedba višestrukih klasifikatora između više skupova podataka još je uvijek teorijski neistražena i ostavljena raznim ad hoc procedurama koje ili imaju manjak statističke pozadine ili koriste statističke metode pogrešno. Jezgra ovog rada je proučavanje statističkih testova koji mogu (ili jesu) biti korišteni za usporedbu dva ili više klasifikatora na više skupova podataka iz različitih domena. Zadatak je pokazati da li su algoritmi statistički značajno različiti i koji algoritmi odstupaju u performansama. Statistička evaluacija eksperimentalnih rezultata je bila važan dio validacije novih algoritama strojnog učenja dugo vremena. No korišteni testovi su bili nevješti i neprovjereni. Dok su procedure za usporedbu dva klasifikatora na jedinstvenom problemu ponuđene čak desetljeće prije, mehanizmi za usporedbu više klasifikatora na jednom ili više testova, još se uvijek koriste polovično i s nezadovoljavajućim pristupima.

Poglavlje 1

Neuravnoteženi skup podataka

1.1 Klasifikacija u uvjetima neuravnoteženih podataka

Najprije ćemo se upoznati sa pojmom klasifikacije u nadziranom okruženju. U strojnom učenju, glavni cilj klasifikacije jest naučiti sustav da predviđa rezultate dotad neviđenih primjera s mogućnosti generalizacije, a sve prema eksplicitnoj informaciji o primjerima i vrijednosti njihove ciljne funkcije. Dakle, imamo n ulaznih primjera $x_1, ..., x_n$ koji su opisani s i atributa $a_1, ..., a_i \in \mathbb{A}$ te poznate vrijednosti izlazne klase $y_j \in \mathbb{C} = \{c_1, ..., c_m\}$. U tom slučaju, algoritam za učenje generira funkciju koja preslikava $\mathbb{A}^i \rightarrow \mathbb{C}$ i naziva se klasifikator.

U klasifikaciji, skup podataka je nebalansiran kada je broj primjera koji predstavlja jednu klasu manji od broja primjera koji predstavlja drugu klasu. Obično je klasa s najmanjim brojem primjera od interesa. U ovom radu, bavit ćemo se podacima sa samo dvije klase, gdje će jedna predstavljati pozitivnu klasu ili manjinu, dok će druga predstavljati negativnu ili većinsku klasu. U tom slučaju će klasifikator biti pristraniji klasi s više primjera, te će pravila za predviđanje većinske klase imati daleko veći značaj u odnosu na mjeru točnosti, dok će pravila za predviđanje manjinske klase biti ignorirana, to jest tretirana kao šum. Stoga su primjeri pozitivne klase obično pogrešno klasificirani. Iskrivljena distribucija sama po sebi nije problem za nadzirano učenje [21], [15], no problem je niz poteškoća koje se javljaju uz nju. Prvi problem je što nebalansirani skup podataka nema dovoljan broj primjera manjinske klase. U [7] autori pokazuju da razina pogreške opada s povećanjem broja primjera manjinske klase. No, povećanje broja primjera manjinske klase je često neizvedivo u problemima iz stvarnog svijeta. Drugi problem je preklapanje klasa. Tada se potiču generalna pravila koja pogrešno klasificiraju primjere općenito, a pogotovo primjere manjinske klase.

Mnoge metode su razvijene za rješavanje ovog problema. Prema [14], pristupi za rješavanje se dijele na vanjske i unutarnje, pristupe koje unose promjene na razini algo-

ritma i metode kojima se intervenira u procesu uzorkovanja podataka. Na razini algoritma, prilagođava se postojeći klasifikator tako da ima većeg utjecaja na pozitivnu klasu. Tu je važno poznavanje klasifikatora kao i domene primjene kako bi razumjeli zašto klasifikator ne radi dobro u uvjetima nebalansiranosti. Na razini skupa podatka, rebalansira se distribucija klasa na način da se ponovljenim uzorkovanjem premješta prostor podataka. Time se izbjegava modificiranje algoritama jer se pokušava smanjiti efekt nebalansiranosti. Pristupi između dva navedena su troškovno-osjetljive metode koje uključuju transformaciju podataka dodajući im težinu i modifikaciju algoritama da prihvate te težine. Minimizacija greške zamjenjuje se tada minimizacijom ukupnog troška, čime se utežnjavanjem grešaka na manjinskim primjerima nastoji riješiti i problem manjinske klase. U ovom radu promatrani su pristupi koji su bazirani na ansamblima, te uključuju kombinaciju ansambla i jedne od prethodno navedenih metoda, ponajviše metode za balansiranje podataka i troškovno-osjetljive metode. Hibridne metode ansambla na razini skupa podataka obično obrade skup podataka prije treniranja svakog klasifikatora. S druge strane, troškovno-osjetljivi ansambli, umjesto modificiranja baznog klasifikatora, minimiziraju trošak pomoću ansambla.

Pitamo se zašto je učenje iz neuravnoteženog skupa podataka teško? Primjerice, 1-NN može netočno klasificirati mnoge slučajeve iz manjinske klase zato što najbliži susjedi pripadaju većinskoj klasi. U situaciji gdje je neravnoteža vrlo visoka, vjerojatnost da je najbliži susjed manjinske klase primjer većinske klase je vrlo visoka, a ocjena pogreške manjinske klase će imati tendenciju visokih vrijednosti, što je neprihvatljivo. Stabla odlučivanja imaju sličan problem. U nazočnosti preklapanja klasa, stablo odlučivanja bi moralo napraviti mnogo testova kako bi razlikovalo dvije klase. Obrezivanje stabla možda neće ublažiti problem. To dovodi do činjenice da obrezivanje uklanja grane koje smatra previše specifičnim, označavajući nove čvorove listove s dominantnom klasom. Dakle, postoji velika vjerojatnost da će većinska klasa biti također i dominantna na tim čvorovima.

1.2 Performanse klasifikatora u nebalansiranim domenama

Kriterij ocjenjivanja je ključan kod performansi klasifikacije i modeliranja klasifikatora. Najjednostavniji način ocjenjivanja performansa je temeljen na matrici konfuzije. U problemima s dvije klase, ona prikazuje rezultate točno i netočno predviđenih primjera za svaku klasu, što je prikazano tablicom 1.1.

Odatle je moguće izvesti razne metrike za mjerenje performansi klasifikatora za učenje kao što su mjera pogreške

$$\text{Err} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1.1}$$

i mjera točnosti

	Positive prediction	Negative predicition
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

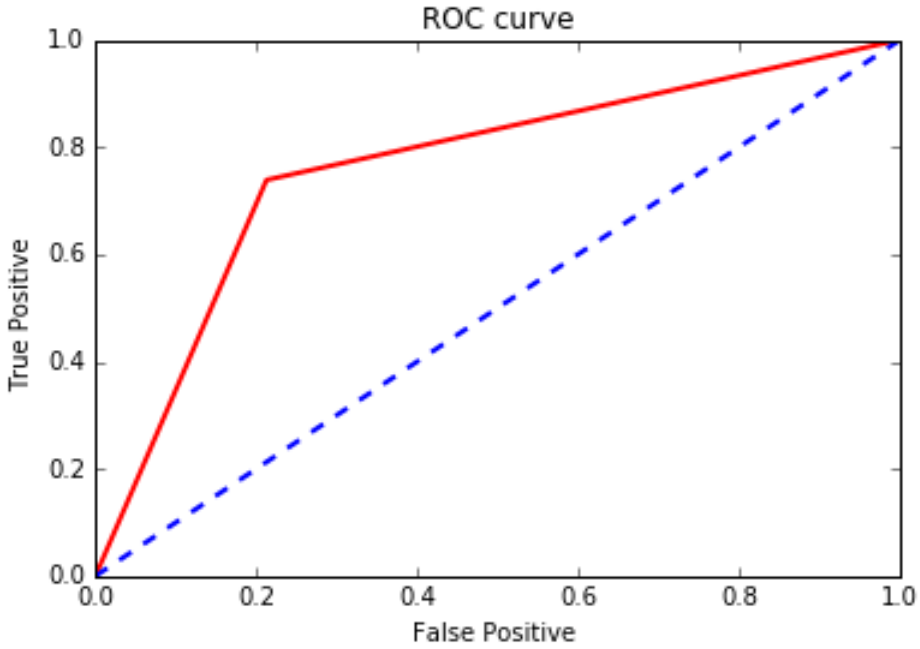
Tablica 1.1: Matrica konfuzije

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} = 1 - \text{Err.} \tag{1.2}$$

Mjera točnosti je najčešće korištena empirijska mjera, no kod nebalansiranog skupa nije najprikladnija jer ne razlikuje dobro točno klasificirane primjere svake klase. Kada su prethodne vjerojatnosti jako različite, korištenje takvih mjera može dovesti do krivih zaključaka. Mjera pogreške i točnosti su posebno upitne za korištenje u takvim slučajevima jer su snažno naklonjene većinskoj klasi pri proučavanju utjecaja distribucije klasa na učenje. Na primjer, jednostavno je kreirati klasifikator koji ima točnost 99% u domeni gdje većinska klasa čini 99% primjera prognozirajući da svaki novi primjer pripada većinskoj klasi. Druga mana mjere točnosti i mjere pogreške jest da te metrike smatraju različite pogreške klasifikacije jednako bitnima. Na primjer, bolesni pacijent koji je dijagnosticiran kao zdrav može biti fatalna pogreška, dok zdrav pacijent koji je dijagnosticiran kao bolestan i nije tako strašna pogreška jer će zahtijevat daljnja istraživanja. Još jedna stvar koje bi trebali biti svjesni prilikom proučavanja utjecaja distribucije klasa na sustave učenja je da se distribucija klasa može mijenjati. Uzmimo primjer matrice konfuzije, mjere koje uzimaju vrijednosti iz obje linije će biti krajnje osjetljive na asimetriju, takve su i mjera točnosti i mjera greške. Kako se distribucija klasa mijenja, mijenjaju se i te mjere, makar se sama izvedba klasifikatora ne mijenja. Tada bi bilo zanimljivije kada bismo koristili pokazatelj koji odvaja pogreške (ili pogotke) koje se pojavljuju u svakoj klasi zasebno. Iz 1.1, možemo izvesti četiri metrike koje mjere izvedbu klasifikacije nezavisno na pozitivnoj i negativnoj klasi. To su:

1. True positive rate: $\text{TP}_{\text{rate}} = \frac{\text{TP}}{\text{TP+FN}}$ broj točno klasificiranih primjera pozitivne klase
2. True negative rate: $\text{TN}_{\text{rate}} = \frac{\text{TN}}{\text{FP+TN}}$ broj točno klasificiranih primjera negativne klase
3. False positive rate: $\text{FP}_{\text{rate}} = \frac{\text{FP}}{\text{FP+TN}}$ broj pogrešno klasificiranih primjera negativne klase
4. False negative rate: $\text{FN}_{\text{rate}} = \frac{\text{FN}}{\text{TP+FN}}$ broj pogrešno klasificiranih primjera pozitivne klase

Cilj klasifikatora je minimizirati FP i FN, tj. maksimizirati TP i TN. Jedan način za kombiniranje ovih mjera jest korištenje ROC krivulje (slika (1.1)) kao kriterija ocjenjivanja.



Slika 1.1: ROC krivulja

ROC (Receiver Operating Characteristic) krivulja (Slika 1.1) dopušta vizualizaciju odnosa između TP (prednosti) i FP (troška) te karakterizira rezultate binarnog klasifikacijskog modela kroz sve granice (treshhold) između TP i FP; pokazuje da niti jedan klasifikator ne može povećati broj točno predviđenih bez da smanji broj pogrešno predviđenih. Ona je konzistentna za dani problem čak i kad je distribucija pozitivne i negativne klase iskrivljena. Svaka granica ima svoju vrijednost na ROC krivulji. Tako spojene točke tvore ROC krivulju. Područje ispod ROC krivulje, AUC (1.3), predstavlja očekivanu izvedbu u jednom broju.

$$\text{AUC} = \frac{1 + \text{TP}_{\text{rate}} - \text{FP}_{\text{rate}}}{2} \tag{1.3}$$

Poglavlje 2

Pred-obrada podataka

2.1 Notacija

Uzimamo notaciju prema [4] kako bi bliže objasnili metode nadziranog učenja. Neka je $E = \{E_1, ..., E_N\}$ skup podataka gdje svaki primjer $E_i \in E$ sadrži oznaku pripadnosti klasi. Svaki $E_i \in E$ je uređeni par $E_i = (\overrightarrow{x_i}, y_i)$ gdje je $\overrightarrow{x_i}$ vektor vrijednosti atributa, a y_i je vrijednost klase. Cilj nadziranog učenja je konstruirati generalni model $y = f(\overrightarrow{x})$, gdje je f nepoznata funkcija, poznata kao konceptualna funkcija koja omogućava predviđanje vrijednosti y za dotad neviđene vrijednosti vektora $\overrightarrow{x}$. Najviše što sustav za učenje može inducirati jest funkcija h koja aproksimira funkciju f, $h(\overrightarrow{x}) \approx f(\overrightarrow{x})$. h se naziva hipoteza konceptualne funkcije f. Tablica 2.1 prema [4] prikazuje skup podataka s N primjera i M atributa.

	A_1	A_2	...	A_M	Y
E_1	x_{11}	x_{12}	...	x_{1M}	y_1
E_2	x_{21}	x_{22}	...	x_{2M}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
E_N	x_{N1}	x_{N2}	...	x_{NM}	y_N

Tablica 2.1: Skup podataka u obliku atributi - vrijednost

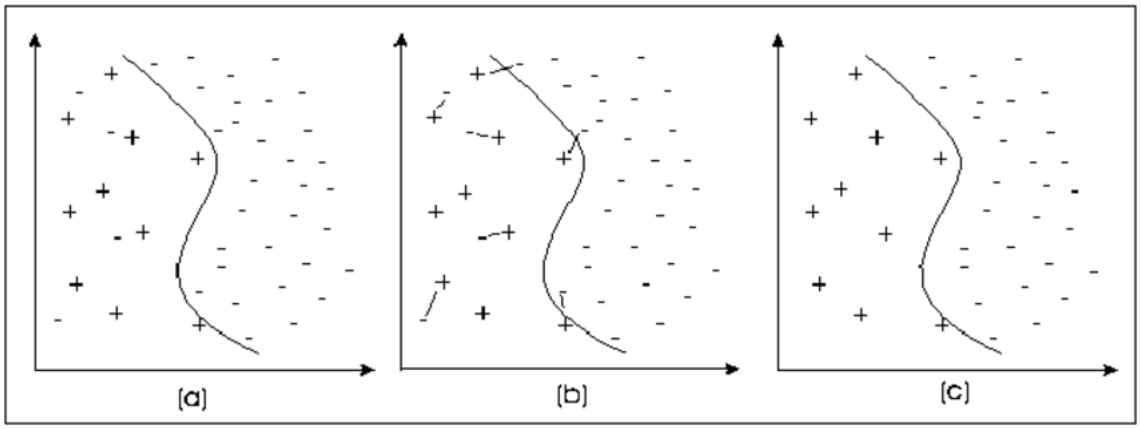
Kolone $(A_1, ..., A_M)$ reprezentiraju atribute, a redci $(E_1, ..., E_N)$ primjere. Vrijednosti x_{ij} predstavljaju vrijednost j-tog atributa za primjer i. U klasifikacijskom problemu, Y predstavlja atribut klase i sadržava skup diskretnih vrijednosti $C = \{C_1, ..., C_{N_{Cl}}\}$. U ovom radu, kao što je već navedeno, proučavat će se problemi dviju klasa, gdje jedna klasa nadmašuje drugu u broju primjera.

2.2 Metode za pred-obradu podataka

Pokazat ćemo nekoliko metoda za balansiranje neuravnoteženog skupa podataka. Ponovno prema [4] bavimo se ne-heurističkim metodama koje služe kao bazne metode. To su slučajne over- i under- sampling metode. Cilj slučajne over-sampling metode je balansirati distribuciju klasa pomoću replikacija primjera manjinske klase, dok slučajne under-sampling metode imaju za cilj balansirati klasu distribucije na način da slučajnom eliminacijom izbacuju primjere većinske klase. Također postoje kombinacije tih metoda. Autori se slažu da slučajne over-sampling metode mogu povećati vjerojatnost da će predobro opisati problem koji sigurno neće vrijediti kao generalno rješenje (overfitting) jer metoda zapravo samo dodaje egzaktne kopije manjinske klase. Problem je što onda klasifikator sadrži pravila koja vrijede samo za konkretan skup podataka. Mana under-sampling metoda je što može izbaciti potencijalno korisne primjere koji mogu biti važni za predikciju. Prednost navedenih metoda je što su nezavisne o osnovnom klasifikatoru.

Tomek links

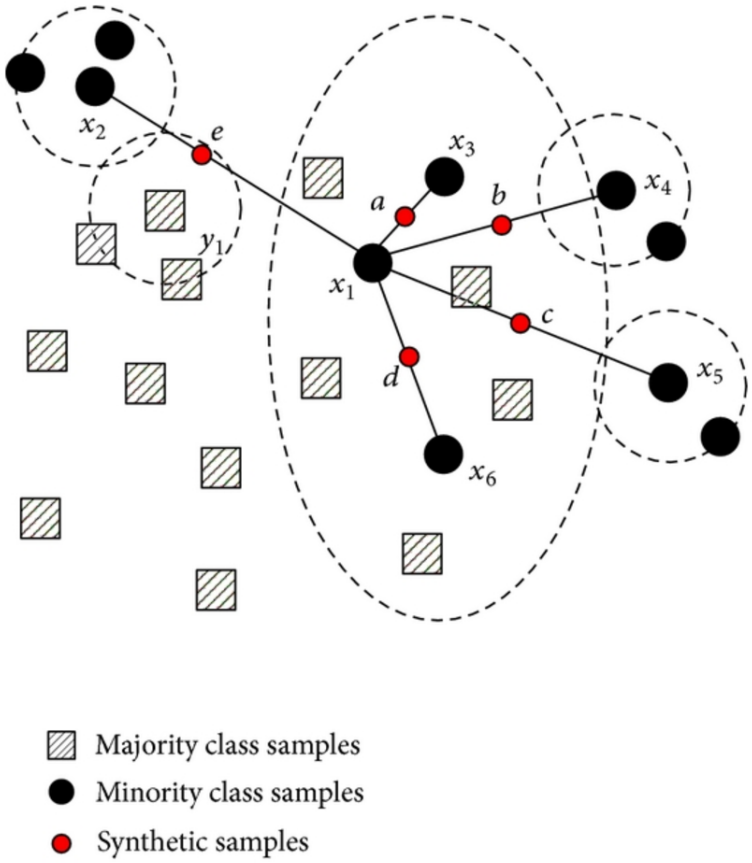
Tomek link je uređeni par (E_i, E_j) za koji vrijedi da ne postoji primjer E_l takav da je $d(E_i, E_l) < d(E_i, E_j)$ ili $d(E_j, E_l) < d(E_i, E_j)$ gdje je d udaljenost između primjera. Ako dva primjera formiraju Tomek link tada je ili jedan šum ili se oba nalaze na rubu. On se koristi kao under-sampling metoda gdje proučavamo samo primjere negativne klase ili kao metoda za čišćenje podataka gdje gledamo cijeli skup podataka. Slika 2.1 prikazuje Tomek links metodu gdje pronalazi primjere koji tvore Tomek link i izbacuje one iz većinske klase.



Slika 2.1: Tomek link

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE je over-sampling metoda čija je glavna ideja kreirati nove primjere manjinske klase interpolacijom među primjerima manjinske klase koji leže zajedno. Ona kreira primjere slučajnim odabirom jednog od k najbližih susjeda i skupa novih primjera dobivenih interpolacijom kao što je prikazano na slici (2.2). Zbog toga je izbjegnuta prekomjerna specijalizacija modela i postavljene su granice odluke da se manjinska klasa proširi u prostor većinske klase.



Slika 2.2: SMOTE

Ova metoda može biti i modificirana tako da razvrstava primjere manjinske klase u tri grupe: sigurni primjeri, primjeri na granici i potencijalni šumovi računajući udaljenosti između primjera. Za sigurne primjere, algoritam slučajnim odabirom izabire jedan od k

najbližih susjeda, za primjere na granici bira najbliži susjed, dok za potencijalne šumove ne radi ništa.

SMOTE + Tomek links

Kombinacijom dviju metoda gdje jedna pripada over-, a druga under- sampling metodi, izbjegavaju se problemi koji ostaju nerazriješeni kad koristimo samo jednu metodu, posebno u slučaju iskrivljene distribucije. Često klasteri nisu dobro definirani ukoliko većinska klasa okupira prostor manjinske, također interpolacijom manjinske klase možemo narušiti klastere većinske klase. U takvoj situaciji može doći do prekomjerne specijalizacije modela. Primjena Tomek link-a na over-sampled skup podataka za treniranje služi kao metoda za čišćenje primjera iz obiju klasa.

Poglavlje 3

Ansambli

3.1 Taksonomija

Glavni cilj ansambla je poboljšati performanse klasifikatora dodajući mu druge klasifikatore te tako kombinirati novi klasifikator koji ima performanse bolje od svakog zasebno. Ideja je konstruirati više klasifikatora na treniranom skupu podataka i njihove predikcije agregirati za nepoznate i nove primjere. Time se poboljšava svojstvo generalizacije: svaki stvara greške, ali jer su različiti, pogrešno klasificirani redci nisu jednaki. Pristranost karakteriziramo kao mjeru sposobnosti generaliziranja sve do testnih podataka, dok je varijanca karakterizirana kao mjera opsega skupa za treniranje na koji je klasifikator osjetljiv [14]. Unatoč neslaganju autora oko ove tematike, različitost klasifikatora je iznimno bitna za formiranje ansambla. Primijetimo ipak da odnos mjera različitosti i točnosti nije prikazan, no više zbog mjere različitosti nego zbog te relacije. Postoji nekoliko načina da postignemo različitost, važno je da bazni klasifikator bude onaj koji slabo uči, to znači da male promjene na skupu podataka dovode do velikih promjena u modelu. Najkorišteniji ansambli za učenje su AdaBoost [11] i Bagging [6] zbog značajnih poboljšanja u klasifikaciji. U tim metodama klasifikatori su strateški generirani da postignu različitost koja im je potrebna, manipulirajući skup za treniranje prije učenja svakog klasifikatora.

Bagging

Breiman [6] uvodi u koncept bootstrap-a za konstruiranje ansambla. On se sastoji od treniranja različitih klasifikatora s bootstrapped replikacijama orginalnog skupa za treniranje. Dakle, novi skup podataka je formiran za treniranje svakog klasifikatora tako da na slučajni način mijenja orginalni skup podataka. Različitost se postiže u ponovnom uzorkovanju na različitim podskupovima podataka. Varijacije Bagging metode za velike skupove podataka, gdje svaki podskup tog velikog skupa služi za treniranje nekog od klasifikatora, su Rvo-

tes, gdje podskup kreira na slučajan način, i Ivotes, gdje su podskupovi uzastopni ovisno o važnosti primjera; važni primjeri su oni koji poboljšavaju različitost. Teški primjeri su detektirani pomoću klasifikatora koji izlaze iz okvira, to jest primjer je težak ukoliko se pogrešno klasificira pomoću klasifikatora ansambla. Ti teški primjeri su uvijek dodavani slijedećem podskupu gdje laki primjeri imaju male šanse da budu uključeni.

Boosting

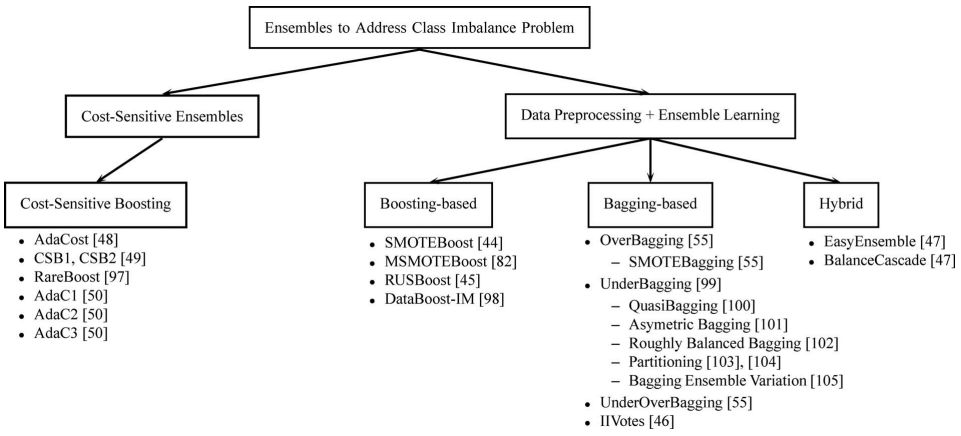
Boosting je prvi uveo Schapire 1990. godine [11] i on je dokazao da se slabi klasifikatori mogu pretvoriti u jake u smislu PAC učenja. PAC učenje kaže [23], ukoliko je neka hipoteza pogrešno izražena, tada će to biti vidljivo već na malom podskupu primjera s velikom vjerojatnošću i obratno. AdaBoost je najreprezentativniji algoritam ove familije i jedan je od najboljih algoritama za rudarenje podacima. On koristi cijeli skup podataka za treniranje svakog klasifikatora, ali u svakom novom koraku stavlja fokus na teške primjere s ciljem točne klasifikacije u slijedećoj iteraciji onih primjera koji su bili pogrešno klasificirani u prethodnoj iteraciji. Fokus na teške primjere se ovdje postiže promjenom težina individualnih primjera kroz iteracije algoritma; u svakoj novoj iteraciji se povećava težina pogrešno klasificiranih vrijednosti i suprotno, smanjuju se težine točno klasificiranih. Također, dodaju se i težine posebno za svaki klasifikator ovisno o njegovoj sveukupnoj točnosti, više značajnosti se dodaje pouzdanijim klasifikatorima. Klasa svakog novog primjera jest klasa u većini ovisno o klasifikaciji svakog klasifikatora.

3.2 Ansambli u uvjetima nebalansiranosti podataka

Kao što je već napomenuto, niti jedan od ovih algoritama ne rješava problem nebalansiranosti klasa direktno, ali posljednjih godina su dani kao potencijalno rješenje. Razlikujemo četiri familije ansambla za ovaj problem. Najprije, tu su troškovno-osjetljivi boosting pristupi koji su slični troškovno-osjetljivim metodama, ali gdje boosting algoritam minimizira trošak. Kod ostalih familija je ugrađena tehnika pred-obrade podataka i to je ono što im je zajedničko, a razlikujemo ih prema ansamblu koji koriste. Tako nalazimo boosting i bagging ansamble te hibridne. Na slici 3.1 prikazano je stablo metoda prema predloženoj taksonomiji u uvjetima nebalansiranosti.

Troškovno-osjetljivi boosting

AdaBoost je algoritam orijentiran prema točnosti i u uvjetima nebalansiranih podataka skloniji je većinskoj klasi jer utječe na sveukupnu točnost. Zbog toga modificiramo algoritam na način da primjeri iz različitih klasa nisu jednako tretirani. Kako bi se to postiglo, troškovno-osjetljiv pristup uzima generalni okvir od AdaBoost-a, ali istovremeno



Slika 3.1: Predložena taksonomija ansambla u uvjetima nebalansiranosti podataka

uključuje trošak u formulu za težine. U jednom od algoritama, AdaCost, težina se ažurira dodavanjem funkcije φ koja je prilagođena trošku. Ta funkcija povećava težine „naviše“ primjerima s većim faktorom troška ukoliko je primjer pogrešno klasificiran.

Boosting ansambli

SMOTEBoost i MSMOTEBoost su metode koje koriste SMOTE i MSMOTE kao algoritme za pred-obradu podataka, respektivno. Težine novih primjera su proporcionalne ukupnom broju primjera u novom skupu podataka, stoga su njihove težine uvijek jednake u svim iteracijama i za svake nove primjere gdje su težine originalnog skupa podataka normalizirane tako da tvore distribuciju s novim primjerima. Nakon treniranja klasifikatora, ažuriraju se težine originalnog skupa podataka i ponovno se uzorkuju novi primjeri. Takvo ponavljanje pridonosi različitosti u treniranju, što je generalna prednost ansambla. RusBoost je algoritam sličan SMOTEBoost, ali koristi undersamling metodu za balansiranje, to jest rješava se primjera većinske klase u svakoj iteraciji. U ovom slučaju nije potrebno označavati nove težine primjerima, jednostavno je dovoljno normalizirati težine preostalih primjera s ukupnom težinom. Ostatak je isti kao i kod SMOTEBoost-a.

Bagging ansambli

Mnogi pristupi su razvijeni koristeći bagging ansamble u uvjetima nebalansiranih podataka zbog njihove jednostavnosti i dobre sposobnosti generalizacije. Integracija procesa pred-obrade je jednostavnija u bagging algoritme nego u boosting. Bagging algoritmi ne

uključuju ponovno računanje težina; u ovim algoritmima je glavni faktor način na koji se prikupljaju replike dobivene pomoću bootstrapa.

(a) OverBagging: jednostavan način kako nadići problem iskrivljenosti klasa je uzeti u obzir klase primjera koje su na slučajni način kreirani iz originalnog skupa podataka. Proces prekomjernog uzorkovanja se može obaviti prije treniranja svakog od klasifikatora. Prekomjerno uzorkovanje uključuje povećanje broja primjera manjinske klase pomoću njihovih replikacija, a svi primjeri većinske klase mogu biti uključeni u novi bootstrap, ili bolja opcija je da ih podijelimo i tako povećamo različitost. Još jedan drugačiji način za prekomjerno uzorkovanje je koristeći SMOTE kao proces pred-obrade. U SMOTEBagging metodi, obje klase doprinose svakom skupu s N_{maj} primjera, ali SMOTE postotak uzorkovanja (a%) je postavljen u svakoj iteraciji (rangiran od 10% do 100% u zadnjoj iteraciji, uvijek pomnožen s 10) i definira broj pozitivnih primjera (a% $\cdot$N$_{maj}$) koje su na slučajan način ponovno uzorkovane iz originalnog skupa podataka. Ostatak pozitivnih primjera je generiran pomoću SMOTE algoritma, dok su negativni primjeri bootstraped u svakoj iteraciji.

(b) UnderBagging: u suprotnosti s OverBagging, koristi undersampling metode pred-obrade. Obično se primjenjuje samo na većinsku klasu; također ponovno uzorkovanje sa zamjenom manjinske klase se također može primijeniti. Primijetimo da je veća vjerojatnost da zanemarimo koristan negativan primjer, ali svaki skup ima manje podataka nego početan skup.

(c) UnderOverbagging ansambli: ovi algoritmi koriste različitu metodologiju od Overbagginga do Underbagging i sličnu SMOTEBaggingu. Koristi metode pod-uzorkovanja i prekomjernog uzorkovanja; postotak uzorkovanja je postavljen u svakoj iteraciji, od 10% do 100%, i označava broj primjera uzet iz svake klase. Stoga, prvi klasifikator je treniran na manjem broju primjera nego zadnji.

Hibridni ansambli

Razlika ovih algoritama od prethodnih je što koriste dvostruko učenje ansambla, to jest kombiniraju i bagging i boosting. Finalni klasifikator je zapravo ansambl ansambla. Oni koriste bagging kao glavni algoritam, ali u svrhu treniranja klasifikatora za svaki novi skup, treniraju svaki skup pomoću AdaBoosta. Kao i underbagging, svaki balansirani skup je konstruirana pomoću slučajnih pod-uzorkovanih primjera iz većinske klase i svih primjera manjinske klase. EasyEnsemble je pristup koji ne vrši nikakve operacije na primjerima originalnog skupa podataka nakon svake iteracije AdaBoosta, stoga svi klasifikatori mogu biti trenirani u paraleli.

Poglavlje 4

Statistike i testovi

Mnoga istraživanja pokazuju da ne postoji utvrđena procedura za usporedbu klasifikatora kroz više skupova podataka. Razni znanstvenici su usvojili različite statistike i tehnike za proučavanje je li razlika između algoritama stvarna ili slučajna. Krećemo od rezultata algoritama na skupovima podataka i procjenjujemo ih koristeći AUC ili neku drugu mjeru. Jedini zahtjev je da kompilirani rezultati pružaju pouzdane procjene učinkovitosti algoritama za svaki skup podataka. Ti brojevi dolaze od ponovljene podjele skupa podatka na testni dio i dio za treniranje. Postoji fundamentalna razlika između testova za usporedbu dva klasifikatora na jednom skupu podataka i na više raznih skupova. Kada testiramo na jednom skupu podataka, obično računamo srednju vrijednost i varijancu na različitim dijelovima skupa podataka za testiranje i treniranje. Jer su ti uzorci obično zavisni, potrebno je mnogo truda za kreiranje statističke procedure kako bi se izbjegla pristranost varijance. Višestruke podijele svakog skupa podataka, služe jedino za procjenu performansi rezultata, a ne i varijance. Izvor varijance su razlike u performansama između zavisnog i nezavisnog dijela uzorka kako pogreška 1. vrste ne bi bila problem. Možemo koristit razne vrste cross-validacije ili metode da jedan dio uzorka izostavimo zbog toga što višestruke podijele nisu pristrane rezultatu procjene. Problem ispravnog statističkog testa za usporedbu klasifikatora na jednom skupu podataka nije srodan usporedbi na više skupova podataka, no usporedba na više skupova podataka prirodno daje uzorak nezavisnih mjera pa je ta usporedba još jednostavnija.

4.1 Usporedba dva klasifikatora

T-test

Najprije upozoravamo protiv prekomjerne upotrebe t-testa kao obično konceptualno neprikladnog i statistički nesigurnog [8]. Proučavamo da li je prosječna razlika u performan-

sama kroz skupove podataka između dva klasifikatora statistički značajno različita od nule. Neka su c_i^1 i c_i^2 rezultati dva klasifikatora na i-tom od N skupova te neka je d_i razlika $c_i^2 - c_i^1$. Statistika t se računa kao $\overline{d}/\sigma_{\overline{d}}$ i ima studentovu distribuciju s $N-1$ stupnjem slobode.

T-test pati od tri slabosti. On ima smisla jedino ako su razlike kroz skupove podataka sumjerljive. U tom slučaju, korištenje t-test-a za usporedbu dva klasifikatora ima malo smisla u usporedbi s računanjem prosjeka kroz skupove podataka. Prosječna razlika $\overline{d}$ jednaka je razlici između prosječnih rezultata dvaju klasifikatora, $\overline{d} = \overline{c}_2 - \overline{c}_1$. Jedina razlika ove forme t-test-a i računanja dvaju prosjeka direktno je u nazivniku. T-test smanjuje standardnu pogrešku $\sigma_{\overline{d}}$ pomoću varijance među skupovima. Slijedeći problem je, ukoliko nam uzorak nije dovoljno velik (više od 30 skupova podataka), što t-test zahtjeva da su razlike normalno distribuirane. Priroda našeg problema ne daje nikakve odredbe normalnosti, a broj skupova podataka je obično manji od 30. Ironično, Kolmogorov-Smirnovljev test za normalnost i slični su slabi testovi na malim uzorcima, nisu u stanju uočiti abnormalnosti i upozoriti na nekorištenje t-test-a. Dakle, za korištenje t-testa potrebna je normalna distribucija zbog malog uzorka, ali mali uzorak sprječava pronalaženje distribucije.

Wilcoxonov test označenih rangova

Wilcoxonov test označenih rangova je ne-parametarski test koji rangira razlike u performansama dva klasifikatora za svaki skup podataka, ignorirajući predznake, i uspoređuje rangove pozitivnih i negativnih razlika. Neka je d_i razlika u rezultatima dvaju klasifikatora na i-tom skupu. Razlike su rangirane obzirom na njihove apsolutne vrijednosti, dok je prosjek uzet ukoliko imamo više jednakih. Neka je R^+ (4.1) suma rangova pozitivnih razlika, dok je R^- (4.2) suma rangova negativnih razlika.

$$R^+ = \sum_{d_i>0} rang(d_i) + \frac{1}{2}\sum_{d_i=0} rang(d_i) \tag{4.1}$$

$$R^- = \sum_{d_i<0} rang(d_i) + \frac{1}{2}\sum_{d_i=0} rang(d_i) \tag{4.2}$$

Neka je T manja od suma, $T = \min(R^+, R^-)$. U mnogoj literaturi generalnih statistika nalazimo kritične vrijednosti za T ovisno o N. Za veliki uzorak, statistika

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{4.3}$$

je distribuirana aproksimativno normalno. S α=0.05, nulta hipoteza može biti odbačena ako je z manje od -1.96.

Wilcoxonov test je osjetljiviji nego t-test, on pretpostavlja sumjerljivost razlika, ali kvalitativno: veće razlike je broje više, što je poželjno, a apsolutne magnitude su ignorirane. Statistički gledano, test je sigurniji ukoliko ne zahtjeva normalnost podataka. Također, ekstremne vrijednosti imaju manje utjecaja na Wilcoxonov test nego na t-test. Wilcoxonov test pretpostavlja kontinuirane razlike, te ne bi trebale biti zaokruživane jer zaokruživanje decimala može oslabiti test zbog previše jednakih vrijednosti.

4.2 Usporedba više klasifikatora

Problem višestrukih hipoteza je vrlo poznati statistički problem. Cilj je kontrolirati familywise error, vjerojatnost da će se dogoditi barem jedna pogreška prve vrste u nekoj od usporedbi.

ANOVA ponovljenih mjerenja

Metoda za testiranje razlika između više uzoraka je ANOVA ponovljenih mjerenja. Testiramo nultu hipotezu da su performanse svih klasifikatora jednake. ANOVA razdvaja ukupnu varijabilnost na varijabilnost između klasifikatora, varijabilnost između skupova podataka i varijabilnost reziduala. Ako je varijabilnost između klasifikatora značajno veća od varijabilnosti reziduala, možemo odbaciti nultu hipotezu i zaključiti da postoje neke razlike između klasifikatora. Pomoću post-hoc testova možemo provjeriti koji od klasifikatora se ističe kao drukčiji. Tu su najpoznatiji Tukey test (Tukey, 1949.) [22] za usporedbu svakog klasifikatora sa svakim, te Dunnett test (Dunnett, 1980.) [10] za usporedbu svih klasifikatora uz kontrolu. Obje procedure računaju standardnu pogrešku razlika između dva klasifikatora tako što dijele varijancu reziduala s brojem skupova podataka. Kako bi napravili usporedbu para klasifikatora, pripadne razlike podijelimo sa standardnom pogreškom i usporedimo s kritičnom vrijednosti. Taj postupak podsjeća na t-test, osim što su Tukey i Dunnett kritične vrijednosti veće kako bi osigurale da postoji 5% šanse da će jedna od razlika biti pogrešno označena kao značajna. ANOVA pretpostavlja nekoliko stvari. Pretpostavlja, najprije, da uzorak dolazi iz normalne distribucije. Drugo i najvažnije, pretpostavlja sfernost, to jest pretpostavlja jednakost varijanci razlika između svih kombinacija. Povreda ovih pretpostavki ima još veći utjecaj na post-hoc testove.

Friedman test

Friedmanov test (Friedman, 1937., 1940) [12], [13] je ne-parametarski test ekvivalentan ANOVA-i s ponovljenim mjerenjima. On rangira algoritme za svaki skup podataka posebno, a u slučaju jednakosti uzima prosjek. Neka je r_i^j rang j-tog algoritma na i-tom skupu

podataka, te neka je k algoritama i N skupova podataka. Friedmanov test uspoređuje prosjek rangova algoritama, $R_j = \frac{1}{N}\sum_i r_i^j$. Pod nultom hipotezom da su svi algoritmi jednaki u performansama, pa i njihovi rangovi trebaju biti jednaki, Friedmanova statistika (4.4) glasi

$$\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}] \tag{4.4}$$

i ima χ_F^2 distribuciju s $k-1$ stupnjem slobode kada su N i k dovoljno veliki ($N > 10, k > 5$), dok se za manji uzorak algoritama i skupova podataka kritične vrijednosti mogu izračunati. Iman i Davenport (1980.) [19] nude bolju statistiku (4.5)

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{4.5}$$

koja ima F distribuciju s $k-1$ i $(k-1)(N-1)$ stupnjeva slobode. Kada odbacimo nultu hipotezu, možemo nastaviti s post-hoc testovima. Nemenyi test (Nemenyi, 1963.) [20] kaže da je razlika između dva klasifikatora značajna ukoliko se pripadni prosjek rangova razlikuje od kritične vrijednosti

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{4.6}$$

gdje su q_α kritične vrijednosti studentove statistike podijeljene s $\sqrt{2}$. Također možemo koristiti Bonferroni korekciju za kontrolu family-wise pogreške u testiranju više hipoteza. Testna statistika za usporedbu i-tog i j-tog klasifikatora koristeći jednu od navedenih metoda je

$$z = (R_i - R_j)/\sqrt{\frac{k(k+1)}{6N}}. \tag{4.7}$$

Pomoću z i prikladne α tražimo odgovarajuću vjerojatnost iz tablice normalne distribucije. Bonferroni-Dunn test (Dunn, 1961.) [9] kontrolira family-wise pogrešku tako da podijeli α s brojem usporedbi. Alternativni način je da izračunamo CD koristeći istu formulu kao i za Nemenyi test uz kritične vrijednosti $\alpha/(k-1)$. Istraživanja [8] pokazuju da post-hoc testovi imaju veću jakost kada su svi klasifikatori uspoređeni s kontrolnim klasifikatorom, a ne međusobno.

Promotrimo sad hipoteze poredane po značajnosti, zapravo poredamo p vrijednosti po veličini. Najjednostavnije metode su Holm (1979.) [17] i Hochberg (1988.) [16] koje uspoređuju svaki p_i s $\alpha/(k-i)$, a razlikuju se po redoslijedu hipoteza. Holm je step-down procedura koja počinje s najznačajnijom p-vrijednosti. Ako je p_1 manja od $\alpha/(k-1)$, onda se pripadna hipoteza odbacuje i uspoređuje p_2 s $\alpha/(k-2)$. Taj postupak se ponavlja sve dok ne dođe do hipoteze koju ne može odbaciti, te zadrži i sve one koje ostaju nakon nje.

Hochberg je step-up procedura koja djeluje obrnuto, uspoređuje najveću p-vrijednost s α, pa slijedeću s $\alpha/2$ dok ne dođe do hipoteze koju može odbaciti, te odbacuje i sve hipoteze nakon nje. Hommel procedura (Hommel, 1988.) [18] je još kompliciranija za računanje i razumijevanje. Traži najveći j za koji $p_{n-j+k} > k\alpha/j$. Ako takav j ne postoji onda procedura odbacuje sve hipoteze, a ako postoji onda odbacuje sve hipoteze za koje vrijedi $p_i \leq \alpha/j$. Holm procedura je jača od Bonferroni-Dunn procedure i nema nikakve pretpostavke na dane hipoteze. Prednost Bonferroni-Dunn procedure je što je jednostavna za opisati i vizualizirati jer koristi isti CD za sve usporedbe. Nadalje, Hochberg i Hommel procedure odbacuju više hipoteza od Holm procedure jer pod nekim okolnostima premašuju opisanu family-wise pogrešku. Ponekad Friedmanov test izvijesti o značajnoj razlici, a onda post-hoc testovi podbace da je otkriju. To govori o slaboj jakosti testa.

4.3 Jakost testa i repliciranje

Kriterij što je zapravo netočno povezan je s odabirom testa, u praksi jedino se može promatrati vjerojatnost odbacivanja nulte hipoteze, što nikad nije povezano s jakosti testa. Prema [8], to se radi na dva načina. Prvo, uzima se razina značajnosti od 5%, što zapravo znači da se promatra ponašanje statistika s p-vrijednostima oko 0.05. Također promatra se prosječna p-vrijednost kao još jedna mjera "jakosti" testa: što su p-vrijednosti manje, vjerojatnije je za test da će odbaciti nultu hipotezu na postavljenoj razini pouzdanosti. Dvije mjere za procjenjivanje jakosti testa dovode do dvije povezane mjere za repliciranje. Bouckaert (2004.) [5] je predložio definiciju koja se može koristiti zajedno s brojem odbacivanja nulte hipoteze. Definirao je repliciranje kao vjerojatnost da dva eksperimenta s istim parom algoritama će dati jednak rezultat, te je osmislio optimalni nepristrani procjenitelj,

$$R(e) = \sum_{1 \leq i < j \leq n} \frac{2I(e_i = e_j)}{n(n-1)}, \tag{4.8}$$

gdje je e_i ishod i-tog eksperimenta (e_i je 1 ako je nulta hipoteza prihvaćena, 0 ako je odbačena) i I je indikatorska funkcija (1 ako je tvrdnja u zagradama istinita, 0 ako je lažna). Također, opisuje jednostavan način za računanje $R(e)$: ako je nulta hipoteza prihvaćena u p slučajeva i odbačena u q od n slučajeva, onda je

$$R(e) = \frac{p(p-1) + q(q-1)}{n(n-1)}. \tag{4.9}$$

Minimalna vrijednost od R je 0.5, kada $p = q = n/2$, a najveća, 1, kada su ili p ili q jednaki nuli. Nedostatak ove mjere je da će statistički test pokazati slabu repliciranost kada je razlika između klasifikatora marginalno značajna. Kada se uspoređuju dva testa različite jakosti, obično će se onaj čiji je rezultat bliži α smatrati manje pouzdanim. Ako

je jakost procijenjena prosjekom p-vrijednosti, repliciranost je prirodno definirana kroz njihove varijance. Varijanca od p je između 0 i 0.25. U usporedbi s Bouckaertovim $R(e)$, Demšar [8] definira repliciranost u ovisnosti o varijanci od p,

$$R(p) = 1 - 2 \cdot \text{var}(p) = 1 - 2\frac{\sum_i (p_i - \overline{p})^2}{n-1}. \tag{4.10}$$

Problem definicije (4.10), kada se koristi u eksperimentima, je, kada pristranost raste, varijabilnost skupa podatka pada, a time i varijanca od p. Veličina efekta ovisi o broju skupova. Jer su navedene definicije repliciranosti povezane, možemo pisati

$$R(e) = \sum_{1 \le i < j \le n} \frac{2(1 - (e_i - e_j)^2)}{n(n-1)}, \tag{4.11}$$

odakle je lako izvesti

$$R(e) = 1 - 2\frac{\sum_i (e_i - \overline{e})^2}{n-1}. \tag{4.12}$$

Poglavlje 5

Eksperimentalna evaluacija

5.1 Algoritmi i parametri

U ovom radu, promatrano je nekoliko ansambla te njihovo ponašanje u različitim domenama nebalansiranih podataka. Prije učenja, korišten je SMOTE algoritam za predobradu podataka, a od ansambla su korišteni razni primjeri bagging i boosting te kombinacije koje su kreirane da rade u uvjetima neuravnoteženog skupa podataka. Primijetimo da želimo analizirati razlike između algoritama u danom problemu te pokazati koji algoritam ima najbolje performanse u svakom od problema. Slabost troškovno-osjetljivih pristupa je u definiciji funkcije troška. Trošak obično nije prikazan u klasifikacijskim skupovima podataka i postavljeni su ad-hoc ili pronađeni među mogućim troškovima. Postavljamo prilagodljivu strategiju troška, gdje je trošak pogrešno klasificiranih primjera manjinske klase uvijek jednak $C_{\min} = 1$, a trošak pogrešno klasificiranih primjera većinske klase je obrnuto proporcionalan s IR. Od boosting metoda koje su namijenjene za rad u uvjetima nebalansiranosti, pokazane su SMOTEBoost i RusBoost, od bagging metoda BalancedBagging kao kombinacije s undersampling metodama, od hibridnih metoda EasyEnsemble te BalancedRandomForest gdje je RandomForest bazni klasifikator. Slika (5.1) prikazuje korištene metode kao kombinaciju procesa pred-obrade i ansambl algoritma.

Korišteni su python-ovi testovi iz paketa scipy.stats [2] za usporedbu rezultata klasifikatora na svim skupovima podataka u odnosu na metriku AUC. Za korištenje post-hoc testova korišten je python-ov paket scikit-posthocs [1].

5.2 Skupovi podataka

Korišteno je 10 binarnih skupova podataka (slika (5.2)) iz različitih domena sa UCI repozitorija za podatke [3] koji je dostupan za javnost. Svaki skup podataka sadrži dvije klase i one su u raznim odnosima neuravnoteženosti.

	metoda	proces pred-obrade	ansambl	algoritam
SB	SMOTEboosting	oversampling (SMOTE)	AdaBoost	
RB	RusBoost	undersampling	AdaBoost	
BB	BalancedBagging	undersampling	bagging	
EE	EasyEnsemble	undersampling	UnderBagging + AdaBoost	
BRF	BalancedRandomForest	undersampling		Random Forest

Slika 5.1: Metode korištene u eksperimentalnom dijelu

dataset	broj atributa	broj redaka	većina	manjina	IR
hcc	49	165	133	32	4,16
electrical_grid	14	10000	6380	3620	1,76
insurance	86	5822	5474	348	15,73
tictactoe	9	958	626	332	1,89
wilt	6	4889	4578	261	17,54
cervical_cancer	36	858	803	55	14,60
relax	13	182	128	52	2,46
parkinson	754	756	564	192	2,94
diabetic	20	1151	1057	94	11,24
advertissements	1558	3279	2820	459	6,14

Slika 5.2: Skupovi podataka korišteni u eksperimentalnom dijelu

Jer niti jedan klasifikator nije optimalan za svaki mogući skup podataka, simuliran je pokus u kojem će se istaknuti prednosti i nedostaci klasifikatora. To je omogućeno računanjem mjera klasifikacije na svim skupovima podataka. Analiza je provedena za mjeru AUC (slika (5.3)). Prije statističke usporedbe, dane su osnovne statistike rezultata iz tablice na slici (5.3) (slika (5.4)).

5.3 Statistička usporedba dva klasifikatora

T-test

Rezultati t-test-a nad razlikama para klasifikatora mogu se pogledati u tablici (5.1).

Statistika t iz tablice (5.1) ima studentovu razdiobu s 9 stupnjeva slobode. Zbog malog uzorka, ne možemo biti sigurni zadovoljavaju li razlike uvjet normalnosti, pa ga provjeravamo Kolmogorov-Smirnovljevim testom i rezultati su dani u tablica (5.2).

KLASIFIKATOR DATASET	BB	BRF	EE	RB	SB
AD	0.922444	0.971125	0.943504	0.923152	0.928955
CCANCER	0.715923	0.913474	0.905936	0.837312	0.922425
DIABETIC	0.474368	0.600137	0.558612	0.516405	0.548530
EGRID	0.999444	0.999682	0.999444	0.999444	0.999444
HCC	0.658333	0.716667	0.700000	0.700000	0.700000
INSURANCE	0.670874	0.697092	0.690089	0.677974	0.575541
PARKINSON	0.763221	0.819290	0.900048	0.772778	0.806945
RELAX	0.506696	0.520089	0.453125	0.517857	0.430804
TICTACTOE	0.863671	0.918579	0.725539	0.743957	0.724657
WILT	0.871249	0.943714	0.969175	0.967506	0.959815

Slika 5.3: Rezultati klasifikatora u odnosu na metriku AUC

		t	pv
BRF	BB	3,56991	0,00603
BRF	EE	1,13359	0,28626
BRF	RB	2,45706	0,03633
BRF	SB	2,37609	0,04149
EE	BB	1,33852	0,21356
EE	RB	1,15298	0,27862
EE	SB	1,81816	0,10240
RB	BB	1,02704	0,33121
RB	SB	0,33828	0,74291
SB	BB	0,47304	0,64745,

Tablica 5.1: Rezultati t-test-a

KLASIFIKATOR	BB	BRF	EE	RB	SB
count	10.000000	10.000000	10.000000	10.000000	10.000000
mean	0.744622	0.809985	0.784547	0.765638	0.759711
std	0.173353	0.167316	0.186801	0.170421	0.196173
min	0.474368	0.520089	0.453125	0.516405	0.430804
25%	0.661468	0.701986	0.692567	0.683481	0.606655
50%	0.739572	0.866382	0.812793	0.758367	0.765801
75%	0.869355	0.937431	0.934112	0.901692	0.927322
max	0.999444	0.999682	0.999444	0.999444	0.999444

Slika 5.4: Osnovne statistike dobivenih rezultata

		s	pv
BB	BRF	0,50010	0,00776
BB	EE	0,44507	0,02528
BB	RB	0,45235	0,02182
BB	SB	0,44472	0,02545
BRF	EE	0,46782	0,01583
BRF	RB	0,49051	0,00965
BRF	SB	0,49358	0,00901
EE	RB	0,47419	0,01382
EE	SB	0,49342	0,00904
RB	SB	0,46609	0,01642

Tablica 5.2: Rezultati Kolmogorov-Smirnovljevog test-a

Ako pogledamo tablicu KS testa, možemo uočiti da su sve p-vrijednosti manje ili jednake 0.05 što nam govori da ćemo odbaciti nultu hipotezu o normalnosti razlika između podataka na razini značajnosti od 5%, dok na razini značajnosti od 1% nećemo odbaciti sve nulte hipoteze te ima smisla provesti t-test na takvim razlikama. No, t-test ne odbacuje nultu hipotezu u takvim slučajevima. T-test odbacuje hipotezu da su klasifikatori statistički značajno različiti.

Wilcoxonov test označenih rangova

Koristeći python-ove funkcije za računanje odnosa između dva klasifikatora, dobivamo slijedeće rezultate.

		T	z	pv	z≤ 1.96
BB	BRF	0	-2,80306	0,00506	TRUE
BB	EE	14,5	-1,32508	0,18514	FALSE
BB	RB	9,5	-1,83473	0,06655	FALSE
BB	SB	23,5	-0,40772	0,68348	FALSE
BRF	EE	14	-1,37605	0,16881	FALSE
BRF	RB	5	-2,29341	0,02182	TRUE
BRF	SB	6	-2,19148	0,02842	TRUE
EE	RB	15,5	-1,22315	0,22127	FALSE
EE	SB	8,5	-1,93666	0,05263	FALSE
RB	SB	24,5	-0,30579	0,75977	FALSE

Tablica 5.3: Rezultati Wilcoxonovog testa označenih rangova

Tablica (5.3) nam daje T i z vrijednosti, te možemo zaključiti prema p-vrijednosti koje razlike su nam statistički značajne ili možemo z usporediti s -1.96. Zaključujemo da postoje statistički značajne razlike između klasifikatora BalancedRandomForest i SMOTEBoosting, RusBoost i BalancedBagging.

5.4 Statistička usporedba više klasifikatora

ANOVA ponovljenih mjerenja

ANOVA kao i t-test pretpostavlja normalnost što je u malom uzorku teško provjeriti. Rezultat testa su vrijednost testne statistike $F = 2.6988$ i p-vrijednost $p = 0.0459$. Na razini značajnosti od 5% možemo odbaciti nultu hipotezu o jednakosti varijanci rezultata klasifikatora. Pomoću Tukey testa uspoređujemo performanse svakog klasifikatora sa svakim. Tukey test (tablica (5.4)) odbacuje rezultate ANOVA-e uz zaključak da ne postoje statistički značajne razlike između klasifikatora.

Friedman test

F statistika ima $F(4, 36)$ distribuciju i za $\alpha = 0.05$ kritična vrijednost iznosi 2.634, uz rezultat testne statistike $F = 17.9683$ i p-vrijednost $p = 0.0013$ također odbacuje nultu hipotezu.

		meandiff	lower	upper	reject
BB	BRF	0,0654	-0,1623	0,293	FALSE
BB	EE	0,0399	-0,1877	0,2676	FALSE
BB	RB	0,021	-0,2066	0,2487	FALSE
BB	SB	0,0151	-0,2126	0,2427	FALSE
BRF	EE	-0,0254	-0,2531	0,2022	FALSE
BRF	RB	-0,0443	-0,272	0,1833	FALSE
BRF	SB	-0,0503	-0,2779	0,1774	FALSE
EE	RB	-0,0189	-0,2466	0,2088	FALSE
EE	SB	-0,0248	-0,2525	0,2028	FALSE
RB	SB	-0,0059	-0,2336	0,2217	FALSE

Tablica 5.4: Rezultati Tukey testa

Nadalje, promatramo rezultate Nemenyi testa za usporedbu kombinacije klasifikatora. Pripadna kritična vrijednost je $CD = 2.728 \cdot \sqrt{0.5} = 1.929$ uz $\alpha = 0.05$. Uspoređena je razlika prosječnih rangova svakog klasifikatora i rezultati su u tablici (5.5).

		d	-1.929$\leq d \leq$ 1.929
BRF	BB	6,55	FALSE
BRF	EE	2,50	FALSE
BRF	RB	3,65	FALSE
BRF	SB	3,80	FALSE
EE	BB	4,05	FALSE
EE	RB	1,15	TRUE
EE	SB	1,30	TRUE
RB	BB	2,90	FALSE
RB	SB	0,15	TRUE
SB	BB	2,75	FALSE

Tablica 5.5: Rezultati Nemenyi testa

Primjećujemo da se BalancedBagging statistički razlikuje od ostalih. Također možemo primijetiti da postoje dvije skupine klasifikatora, u jednoj su BalancedBagging i BalancedRandomForest dok su ostale u drugoj. Bonferroni-Dunn test uz kritičnu vrijednost $CD = 2.498 \cdot \sqrt{0.5} = 1.766$ daje slijedeće rezultate (tablica (5.6)) koji su gotovo jednaki rezultatima Nemenyi testa.

Prije korištenja ostalih post-hoc testova, računamo z testnu statistiku za usporedbu i-tog i j-tog klasifikatora i rezultati su dani u tablici (5.7).

		d	-1.766≤ d ≤ 1.766
BRF	BB	6,55	FALSE
BRF	EE	2,50	FALSE
BRF	RB	3,65	FALSE
BRF	SB	3,80	FALSE
EE	BB	4,05	FALSE
EE	RB	1,15	TRUE
EE	SB	1,30	TRUE
RB	BB	2,90	FALSE
RB	SB	0,15	TRUE
SB	BB	2,75	FALSE

Tablica 5.6: Rezultati Bonferroni-Dunn testa

		z	pv
BRF	BB	9,263099	1,99E-20
BRF	EE	3,535534	0,000406952
BRF	RB	5,16188	2,44E-07
BRF	SB	5,374012	7,70E-08
EE	BB	5,727565	1,02E-08
EE	RB	1,626346	0,103876157
EE	SB	1,838478	0,065992055
RB	BB	4,101219	4,11E-05
RB	SB	0,212132	0,832004029
SB	BB	3,889087	0,000100622

Tablica 5.7: Testna statistika i pripadna p-vrijednost za usporedbu dva klasifikatora

Jakost post-hoc testova je veća ukoliko su svi klasifikatori uspoređeni s kontrolnim. Uzet je BalancedRandomForest klasifikator kao kontrolni. Tablica (5.8) daje p-vrijednosti usporedbe baznog klasifikatora s ostalima poredanih po veličini.

		z	pv
BRF	BB	9,263099	1,99E-20
BRF	SB	5,374012	7,70E-08
BRF	RB	5,16188	2,44E-07
BRF	EE	3,535534	0,000406952

Tablica 5.8: Usporedba BalancedRandomForest s ostalima

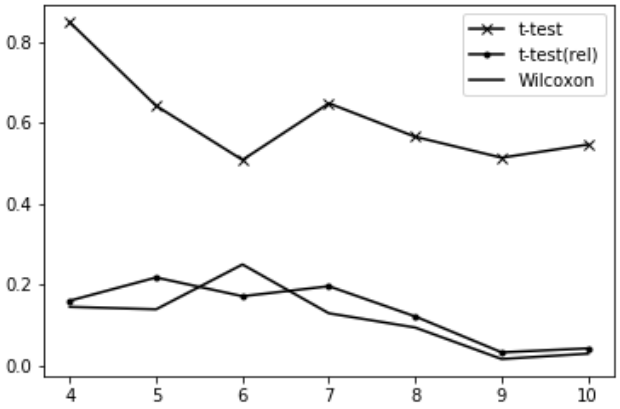
Holm i Hochberg odbacuju sve hipoteze. Slijedeće tražimo najveći j tako da vrijedi Hommel-ov uvjet, no takav j ne postoji pa ponovno odbacujemo sve hipoteze po uvjetu. Zaključujemo da sve procedure govore da postoji statistički značajna razlika klasifikatora od BalancedRandonForest klasifikatora.

5.5 Grafička vizualizacija rezultata

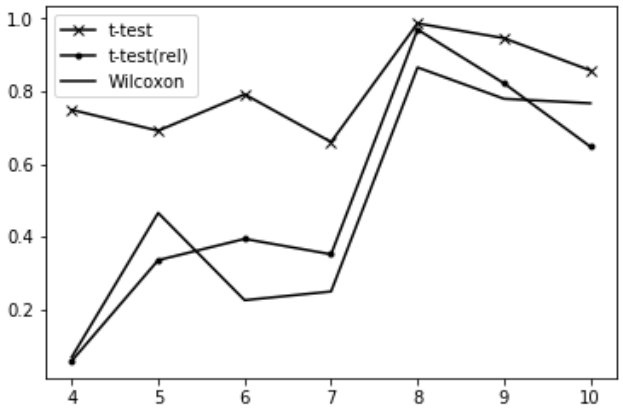
Provjerene su tri statistike za usporedbu dva klasifikatora: t-test na apsolutnim i relativnim razlikama, te Wilcoxonov test označenih rangova. Na slici (5.8) su prikazane prosječne p-vrijednosti dobivene testovima za usporedbu dva klasifikatora, uzete su kombinacije BalancedRandomForest i SMOTEBoost, BalancedBagging i EasyEnsemble te SMOTEBoost i BalancedBagging. Primjetimo da t-test na relativnim razlikama klasifikatora daje slične rezultate kao i Wilcoxonov test, dok je t-test na apsolutnim razlikama najslabiji. Wilcoxonov test daje manje p-vrijednosti pa je veća vjerojatnost da odbaci nultu hipotezu. Poznato je da parametarski testovi odbacuju nultu hipotezu u više slučajeve nego ne-parametarski testovi osim ako nisu zadovoljene pretpostavke testa. Stoga se preporuča korištenje Wilcoxonovog testa ukoliko nisu zadovoljene pretpostavke t-testa.

Za usporedbu više klasifikatora, izračunati su rezultati testova na uzorcima skupova podataka. Rezultati su prikazani na slici (5.9), ponovno na y-osi su dane p-vrijednosti u ovisnosti o pristranosti k. Ne-parametarski Friedmanov test je prikazan kao jači nego parametarski, ANOVA, dok s povećanjem pristranosti, testovi postaju slični. Ponovno je ne-parametarski test jači od parametarskog što je i očekivano uz praktične i teorijske prednosti Friedmanovog testa.

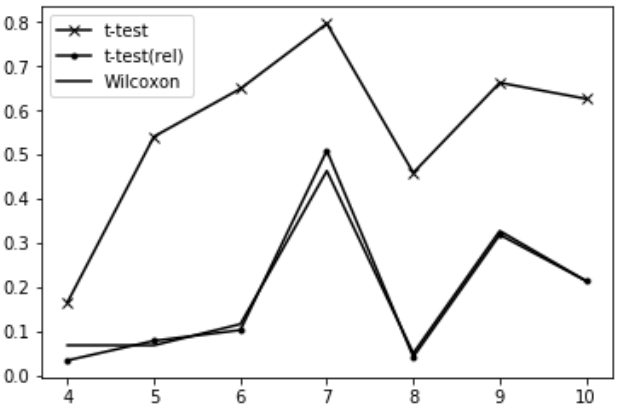
Važno je napomenuti da se radi o relativno malom skupu podataka, pa su rezultati slabije statistički značajni.



Slika 5.5: BalancedRandomForest vs. SMOTEBoost

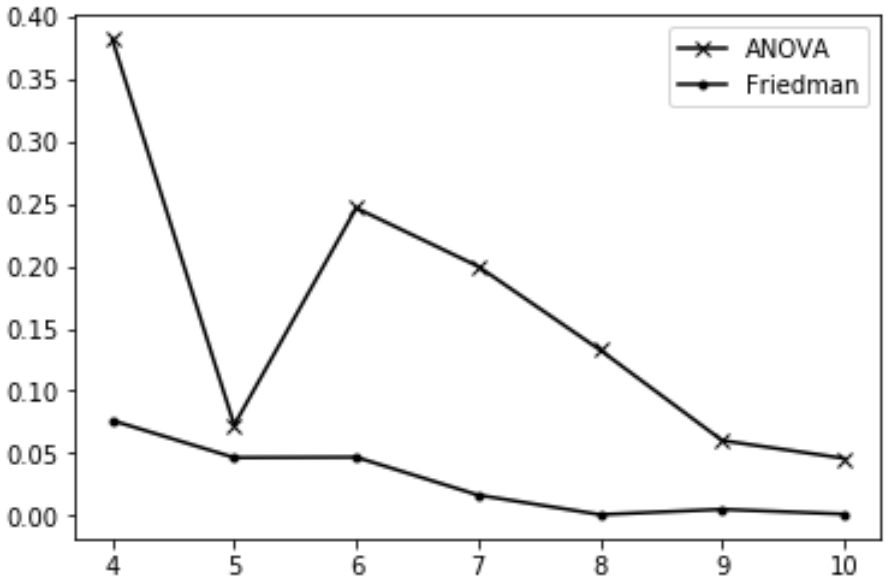


Slika 5.6: SMOTEBoost vs. BalancedBagging



Slika 5.7: EasyEnsemble vs. BalancedBagging

Slika 5.8: Jakost statističkog testa za usporedbu dva klasifikatora: p-vrijednosti kao funkcija pristranosti



Slika 5.9: Jakost statističkog testa za usporedbu više klasifikatora

Bibliografija

[1] *scikit-posthocs Documentation*, `https://buildmedia.readthedocs.org/media/pdf/scikit-posthocs/latest/scikit-posthocs.pdf`.

[2] *SciPy.org*, `http://docs.scipy.org/doc/scipy/reference/stats.html`.

[3] *UCI Machine Learning Repository*, `http://archive.ics.uci.edu/ml/datasets.php`.

[4] G. E. A. P. A. Batista, R. C. Prati i M. C. Monard, *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*, (2004.).

[5] R. R. Bouckaert, *Estimating replicability of classifier learning experiments*, Machine Learning, Proceedings of the Twenty-First International Conference (ICML 2004). AAAI Press (2004.).

[6] L. Breiman, *Bagging predictors*, Mach. Learn. **24** (1996.), 123–140.

[7] N. V. Chawla, *Data mining for imbalanced datasets: An overview*, Data Mining and Knowledge Discovery Handbook (2010.), 875–886.

[8] J. Demsar, *Statistical Comparisons of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research 7 (2006.).

[9] O. J. Dunn, *Multiple comparisons among means*, Journal of American Statistical Association (1961.), br. 56, 52–64.

[10] C. W. Dunnett, *A multiple comparison procedure for comparing several treatments with a control*, Journal of American Statistical Association (1980.), br. 50, 1096–1121.

[11] Y. Freund i R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, J. Comput. Syst. Sci. **55** (1997.), br. 1, 119–139.

[12] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association **32** (1937.), 675–701.

[13] ———, *A comparison of alternative tests of significance for the problem of m rankings*, Annals of Mathematical Statistics **11** (1940.), 86–92.

[14] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince i F. Herrera, *A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches*, IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (2011.).

[15] H. He i E. A. Garcia, *Learning from imbalanced data*, IEEE Trans. Knowl. Data Eng. **21** (2009.), br. 9, 1263–1284.

[16] Y. Hochberg, *A sharper Bonferroni procedure for multiple tests of significance*, Biometrika (1988.), br. 75, 800–803.

[17] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics (1979.), br. 6, 65–70.

[18] G. Hommel, *A stagewise rejective multiple test procedure based on a modified Bonferroni test*, Biometrika (1988.), br. 75, 383–386.

[19] R. L. Iman i J. M. Davenport, *Approximations of the critical region of the Friedman statistic*, Communications in Statistics (1980.), 571–595.

[20] P. B. Nemenyi, *Distribution-free multiple comparisons*, PhD thesis, Princeton University (1963.).

[21] Y. Sun, A. C. Wong i M. S. Kamel, *Classification of imbalanced data: A review*, Int. J. Pattern Recogn. **23** (2009.), br. 4, 687–719.

[22] J. W. Tukey, *Comparing individual means in the analysis of variance.*, Biometrics (1949.), br. 5, 99–114.

[23] T. Šmuc, *Teorija učenja*, Strojno učenje, matrijali s predavanja 2018./2019., `https://web.math.pmf.unizg.hr/nastava/su/materijali/`, 11–43.

Sažetak

Pretpostavka u problemima klasifikacije jest da je broj primjera među različitim klasama podjednak, što nije slučaj u problemima iz stvarnog svijeta. Taj problem je poznat kao učenje u uvjetima neuravnoteženih podataka. Učenje u takvim uvjetima nailazi na mnoge poteškoće, poput pristranosti većinskoj klasi dok je manjinska klasa od interesa. Tu su važne metode za balansiranje podataka kako bi smanjili efekt iskrivljenosti distribucije klasa. Ovdje promatramo metode prekomjernog i pod- uzorkovanja. Metode prekomjernog uzorkovanja mogu bolje inducirati klasifikatore koji su precizniji nego oni inducirani metodom pod-uzorkovanja. Te metode kombiniramo s ansamblima, kao bagging i boosting. Modifikacija ansambla uključuje tehnike za balansiranje prije učenja klasifikatora i ovdje su promatrane razne kombinacije s ciljem pronalaska najboljeg klasifikatora. U nastavku, u ovom radu provodimo eksperimentalnu evaluaciju koja uključuje pet metoda za rad s nebalansiranim podacima na deset skupova podataka s UCI repozitorija. Dalje se bavimo statističkom validacijom rezultata, te proučavanjem da li algoritmi odstupaju u performansama u odnosu na ostale.
Ključne riječi: neuravnoteženi podaci, prekomjerno uzorkovanje, pod-uzorkovanje, ansambli, bagging, boosting, statistička validacija

Summary

Assumption in classification problems is that number of examples in different classes is approximately equal, which is not case in real world problems. The problem is known as learning in conditions of imbalanced data. Learning in such conditions encounters many difficulties, such as bias to the majority class while minority class is interested one. There are important methods for balancing data to minimize skewness of class distribution. Here are observed over-sampling and under-sampling methods. Over-sampling methods can better induce classifier which are more accurate than those induced by under-sampling methods. These methods are combined with ensembles, such as bagging and boosting. Ensemble modification includes balancing techniques before learning the classifiers and here are given different combinations with best classifier purpose. In addition, in this work we perform a board experimental evaluation involving five methods to deal with class imbalanced problems in ten UCI datasets. We are further concerned on statistical validation of results and by examining whether algorithms differ in performance compared to others.
Key words: imbalanced data, over-sampling, under-sampling, ensembles, bagging, boosting, statistical validation

Životopis

Rođena sam 29.03.1995. godine u Karlovcu. Nakon osnovnoškolskog obrazovanja, 2010. godine, upisujem Prirodoslovno-matematičku gimnaziju u Karlovcu. S uspješno položenom maturom, svoje školovanje nastavljam na Prirodoslovno-matematičkom fakultetu u Zagrebu, gdje 2014. godine upisujem preddiplomski studij Matematika. Godine 2017. upisujem diplomski studij Matematička statistika na Prirodoslovno-matematičkom fakultetu, sveučilišta u Zagrebu. Na drugoj godini diplomskog studija polazim Zaba Future Academy u Zagrebačkoj banci u Zagrebu, nakon čega ondje ostajem na studentskom poslu, te 2019. godine potpisujem ugovor i postajem punopravni zaposlenik Zagrebačke banke u zvanju mlađe specijalistice za Poslovnu inteligenciju, IT sektor.