SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Metoda ispitivanja statističke značajnosti u obradi prirodnog jezika

Filip Boltužić
Voditelj: *Prof. dr. sc. Bojana Dalbelo-Bašić*

Zagreb, siječanj 2014.

SADRŽAJ

1. Uvod

Dokazivanje statističke značajnosti neizostavan je dio svake znanstvene publikacije, pa tako i publikacija u području obrade prirodnog jezika (engl. *natural language processing – NLP*). Neki od problema u sklopu obrada prirodnog jezika su klasifikacija govornih činova (engl. *speech acts*) (Pratt, 1977), izgradnja stabla međuovisnosti (engl. *dependency tree*) (Collins, 2003), sažimanje teksta (engl. *text summarizing*), ekstrakcija ključnih riječi (engl. *keyword extraction*) i drugi. Dobivene rezultate za, primjerice, ekstrakciju ključnih riječi potrebno je verificirati. Nije dovoljno riječima argumentirati u korist rezultata već, pošto je postala ustaljena praksa u znanstvenoj zajednici, je potrebno provesti verifikaciju testiranjem statističkih hipoteza. Korištenje statističkih testova u istraživanju standardizira provođenje eksperimenata i daje dodatnu informaciju o rezultatima eksperimenta.

Dobro provedeno istraživanje mora biti potkrijepljeno kvalitetnim statističkim analizama rezultata. Kvalitetna analiza podrazumijeva da pružanje dokaza o tome da su eksperimenti predstavljeni u radu dobiveni nad reprezentativnim podatcima, da nije bilo utjecaja

U obradi prirodnog jezika verifikacija izgrađenog sustava (za prevođenje, klasifikaciju govornih činova, izgradnje stabla međuovisnosti . . .) nužan je korak. Uz argumentirano obrazloženje autora sustava o dobivenim performansama sustava, potrebno je priložiti statistički dokaz kojime se nepobitno pokazuje kako je uistinu ostvaren doprinos znanstvenoj zajednici. Ispitivivanjem statističke značajnosti pokazuje se koliko su rezultati eksperimenata vrijedni, jesu li dobiveni slučajno i u kolikoj mjeri su pouzdani. Ovaj korak je sastavni dio velikog broja NLP članaka.

(Chinchor, 1992) se bavi analizom rezultata MUC-4 (engl. *Message Understanding Conference*), (Koehn, 2004) i (Zhang et al., 2004) ispituju značajnost rezultata strojnog prevođenja (engl. *machine translation*), (Bisani i Ney, 2004) analiziraju rezultate automatskog prepoznavanja govora (engl. *automated speech recognition*). (Berg-Kirkpatrick et al., 2012), (Yeh, 2000) , (Thompson, 1993) se nisu usredotočili na specifičnu metriku (kao što je F1-mjera), već općenitijim metodama ispitivanja statističke

pouzdanosti.

U ovom radu prvo će se objasniti osnove ispitivanja statističke značajnosti. Nakon toga pričat će se o mogućim testovima značajnosti u području strojnog učenja. Svaki spomenuti test bit će kratko objašnjen, s pružanjem prikladnih referenci. Objasnit će se mehanizam pojedinih statističkih testova te prikladnost ili neprikladnost statističkog testa u kontekstu specifičnog problema. Spomenuti problemi bit će tipični problemi iz područja strojnog učenja, s blagim naglaskom na obradu prirodnog jezika. Konačan cilj rada je dati čitatelju uvid u aktualne statističke testove nad problemima strojnog učenja, kako bi što bolje odabrao metodologiju ispitivanja statističku značajnost vlastitih eksperimenata.

2. Testiranje hipoteza

Testiranje hipoteza dominantan je način verifikacije rezultata prilikom objavljivanja znanstvenih radova. Ronald Fisher, otac moderne statistike, pokazao je kako je moguće rezultate eksperimenta dokazati ili opovrgnuti koristeći statističke postupke (Fisher, 1922). Prema Fisherovom testiranju hipoteza moguće je ispitati jesu li rezultati dobiveni eksperimentom statistički značajni ili nisu. Razvojem statistike utemeljen je uobičajen način ispitivanja statističke značajnosti prema kojem se:

1. postavlja inicijalna hipoteza istraživanja,
2. definira nulta H_0 i alternativna hipoteza H_1 (engl. *null and alternative hypothesis*),
3. promatraju statistička obilježja podataka nad kojima će se provoditi odgovarajući statistički testovi (engl. *statistical tests*),
4. na temelju rezultata prethodnog koraka odabire odgovarajući statistički test,
5. odabire relevantna statistička mjera T (engl. *test statistic*),
6. računa distribucija odabrane statističke mjere T pod pretpostavkom da je nulta hipoteza zadovoljena (ove vrijednosti često su unaprijed izračunate i pohranjene u tablicama (Wilcoxon et al., 1973))
7. odabire razina značajnosti α (engl. *significance level*), razina vjerojatnosti ispod koje se odbacuje nulta hipoteza (česti odabiri su 1% ili 5%, ovisno o eksperimentu),
8. računaju granične vrijednosti (engl. *critical region*) distribucije statističke mjere T za razinu značajnosti α,
9. računa statistička mjera t_{obs} dobivena iz eksperimentalnih (stvarnih) podataka,

10. donosi odluka o odbacivanju nulte hipoteze ukoliko je dobivena vrijednost t_{obs} unutar graničnih vrijednosti.

Moguć je i alternativan scenarij prema kojem se na temelju dobivene t_{obs} vrijednosti računa p vrijednost (engl. *p-value*), vjerojatnost eksperimentalnih podataka pod pretpostavkom nulte hipoteze donosi odluka o odbacivanju nulte hipoteze.

Eksperiment kojim je Fisher predstavio testiranje hipotezi je damin test kušanja čaja (engl. *lady tasting tea*). Testom se pokušalo ustanoviti može li dama (u Fisherovom slučaju Muriel Bristol) temeljem okusa razlikovati čaj s mlijekom prema načinu spravljanja napitka (prvo čaj, zatim mlijeko ili obratno). U ovom slučaju nulta hipoteza je pretpostavka da dama ne može razlikovati čaj s mlijekom sudeći samo prema okusu. Dami je predstavljeno osam šalica čaja s mlijekom za koje je morala opisati način pripreme. Dobiveni rezultati uspoređuju se s pravim vrijednostima, te se na temelju toga provodi statistička verifikacija rezultata. Eksperiment je detaljno objašnjen u Fisherovom radu Fisher (1935). Test se smatra izuzetno bitnim za razvoj polja statistike (Potter, 2001).

Interpretacija rezultata statističkog testiranja često se provodi kroz izračunatu p-vrijednost. Na temelju p-vrijednosti, vjerojatnosti $P(H_0|D)$ dobivenih podataka D pod pretpostavkom nulte hipoteze H_0 , donosi se odluka o prihvaćanju ili odbacivanju nulte hipoteze. U slučaju kada je p-vrijednost manja od postavljene razine statističke značajnosti α, nije moguće odbaciti nultu hipotezu zbog nedovoljno dokaza. *P-vrijednost* naziva se i empirijska razina značajnosti.

Testiranje hipoteza primjenjuje se u gotovo svim znanstvenim disciplinama. Nažalost, postupak dokazivanja statističke značajnosti izuzetno je kontroverzno i podložno brojnim kritikama. Objavljen je izuzetno velik broj radova koji kritiziraju provedbe sumnjivih statističkih postupaka, kao što su (Hedges et al., 1985), (Dar et al., 1994), (Yoccoz, 1991). Najprodavanija knjiga iz statistike *How to lie with statistics* (Huff, 2010) pokušava na način pristupačan široj publici pokazati na koji je način moguće slučajno ili namjerno iskoristiti statistiku na pogrešan način. Razlozi pogrešnog provođenja statističkog testa su pogrešno tumačenje ili preskakanje jednog od koraka navedenih na početku poglavlja.

3. Kritike ispitivanje statističke značajnosti

Bruce Thompson (Thompson, 1993) navodi neke kritike konvencionalnih metoda statističkog testiranja:

1. Nulta hipoteza će se **uvijek** odbaciti, ako uzme u obzir dovoljno velika populacija. Thompson kaže:

> Ispitivanje statističke značajnosti može biti vođeno tautologijom. Umorni istraživači, nakon prikupljanja skupa podataka, rade statističke testove kako bi se uvjerili da je prikupljena dovoljno velika količina podataka, što je podatak koji već znaju. Ovakva tautologija učinila je mnogo štete znanstvenoj zajednici.

2. Korištenje ANOVA-e može dovesti do pogrešnih usporedbi. U višedimenzionalnim analizama primjenom hijerarhijskog pristupa moguće je povlačiti usporedbe između podataka iz različitih dimenzija, analogno poslovici: usporedba krušaka i jabuka.

3. Oslanjanje na ispitivanje statističke značajnosti stvara neizbježne dvojbe. ANOVA zahtjeva spajanje varijanci prilikom izračuna srednje devijacije (u nazivniku). Ova operacija je dozvoljena samo u slučaju da su varijable homogene, međusobno usporedive. Slično tome, ANCOVA (analiza kovarijance) pretpostavlja da je zadovoljen uvjet *homogenosti regresije* (engl. *Homogeneity of Regression Slopes*).

3.0.1. Bootstrap

Bootstrap procjenjuje *p-vrijednost*, vadi testne uzorke x_i iz populacije i broji koliko često sustav *A* funkcionira s performansama $\delta(x)$ (ili većim) od sustava *B*. Na raspolaganju stoji samo populacija x, pa se podaci iz x uzorkuju sa zamjenom (engl. *sampling with replacement*). Tako dobiveni uzorci nazivaju se (engl. *bootstrap*) uzorcima.

Za dobivene uzorke x_i bi, prema nultoj hipotezi, trebalo vrijediti da

$$\frac{1}{k}\sum_{i=0}^{k}\delta(x_i) = \delta(x) \tag{3.1}$$

gdje je *k* broj uzoraka. Ako se želi provjeriti može li se odbaciti nulta hipoteza, potrebno je provjeriti *koliko često sustav A daje rezultate koji su bolji od očekivanih*. Očekivani rezultat je da je sustav *A* bolji od sustava *B* za $\delta(x)$. Prema tome, prebrojava se u koliko slučajeva test skupova podataka x_i je *A* bio bolji od *B* za $2\delta(x)$. Pseudokod bootstrap postupka je prikazan je (Berg-Kirkpatrick et al., 2012) bootstrap postupka prikazan je **??**.

Najveća prednost bootstrap metode je mogućnost računanja $\delta(x)$ za bilo koju metriku.

4. Primjeri i tehnike ispitivanja značajnosti u strojnom učenju

Ispitivanje statističke značajnosti rezultata često se obavlja u kontekstu strojnog učenja. Izrada novih tehnika i metoda strojnog učenja zahtjeva dokaz i statističku analizu o performansama sustava.

4.1. Mjerenje performansi modela strojnog učenja

Primjerice, radi se na automatskom prepoznavanju ključnih riječi u tekstu pomoću modela baziranom na strojnom učenju, klasifikatoru. Konačan cilj istraživanja je usporedba i verifikacija performansi dva klasifikatora ključnih riječi, kako bi se pronašao optimalan klasifikacijski sustav. Ulaz klasifikatoru je dokument, a izlaz klasifikatora mora biti niz ključnih riječi. Postoje referentne vrijednosti (engl. *gold standard*), koje se smatraju ispravnim ključnim riječima pojedinog dokumenta s kojima se uspoređuje rezultat dobiven klasifikatorom. U tablici 4.1 dan je primjer ispitivanja rezultata klasifikatora. U prvom stupcu je popis riječi nad kojima se provodi klasifikacija, u drugom stupcu označena je (*T* je točno, *N* netočno) ispravna oznaka riječi, tj. je li neka riječ ključna (*T*) ili nije (*F*). Treći stupac predstavlja izlaz klasifikatora za riječ u odgovarajućem retku. Iz navedene tablice možemo vidjeti kako su moguće četiri kombinacije podudaranja izlaza klasifikatora i stvarne vrijednosti.

Popisivanje svih rezultata može biti nepregledno, pogotovo kod većeg broja testnih primjera, stoga se koriste tablice kontigencije (engl. *contigency tables*). Njima se

Riječ	lopata	život	ja	ljudi	cvijet	ili	svijet	klasifikacija	ritam	latica
Ispravno	T	F	F	T	F	F	T	T	T	F
Dobiveno	T	T	T	F	F	F	T	T	T	T

Tablica 4.1: Rezultat klasifikatora

		Ispravno	
		T	F
Dobiveno	T	TP	FP
	F	FN	TN

Tablica 4.2: Oblik tablice kontigencije

		Ispravno	
		T	F
Dobiveno	T	4	3
	F	1	2

Tablica 4.3: Primjer tablice kontigencije

dobivaju agregirane vrijednosti ispitivanja klasifikatora, a daju pregled koliko ima

- ispravno klasificiranih pozitivnih primjera (engl. *true positives*) –TP,
- ispravno klasificiranih negativnih primjera (engl. *true negatives*) – TN,
- pozitivnih primjera pogrešno svrstanih u negativne (engl. *false positives*) – FP i
- negativnih primjera pogrešno svrstanih u pozitivne (engl. *false negatives*) – FN.

Upravo su ovo sve četiri moguće kombinacije ispitivanja rezultata klasifikatora. Općenit oblik tablice kontigencije prikazan je na tablici 4.2. Tablicu kontigencije za primjer iz tablice 4.1 moguće je vidjeti u tablici 4.3. Ovakav oblik tablica kontigencija često se koristi u strojnom učenju, kao u radu Hall (1999).

Tablica kontigencije predstavlja polaznu točku nekim naprednijim mjerama performansi, a najčešća korištena mjera je preciznost (engl. *accuracy*) *ACC*, računa se

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

.

U praksi, koriste se i mnoga druga mjerila izvedbe klasifikatora, mnoga opisana u radu Powers (2011). Mjere često ovise i o prirodi samog zadatka. Primjerice, strojni prevoditelji rangiraju se prema *BLEU* rezultatu, opisanom u (Papineni et al., 2002). U ovome radu spominjat će se neke druge mjere, no neće se eksplicitno objašnjavati.

4.2. Metodologija ispitivanja značajnosti klasifikatora

Istraživanje novog optimalnog algoritma metode za zadatak strojnog učenja često uključuje izgradnju više modela i promatranje ponašanja tih modela nad različitim skupovima podataka. Ciljevi takvih istraživanja mogu biti: kako pronaći klasifikator koji u prosjeku radi vrlo dobro nad svim dostupnim skupovima podataka, koji klasifikator pruža najbolje rezultate nad malim (specifičnim) skupovima podataka, kada koristiti kombinaciju klasifikatora i sl. Ukoliko se radi o novom zadatku, ne postoji referentni sustav s kojim bi se novoizgrađeni sustav mogao usporediti. U slučaju napadanja poznatog,

donekle rješenog problema potrebno je usporediti perfomanse novodobivenih postupaka s trenutno najboljim. Također, moguće je pretpostaviti da je na raspolaganju jedan ili više skupova podataka. Ti skupovi podataka možda dolaze iz različite domene. Na primjer, izdvajanje ključnih riječ iz novinskih članaka ili proze.

U najjednostavnijem slučaju izrade klasifikatora bi sadržavao:

istraživačku hipotezu,

Nova metoda A radi bolje od trenutno općeprihvaćene metode B. Viši iznos vrijednosti znači da metoda radi bolje.

nultu hipotezu H_0,

Ne postoji razlika u performansama između sustava *A* i *B*.

skup podataka,

Postoji jedinstven skup podataka (engl. *dataset*) – populacija. Moguće je uzimati uzorke x iz populacije.

način mjerenja

c_A je rezultat sustava A, c_B je rezultat sustava B. $d_{A,B} = c_A - c_B$ je razlika u perfomansama između sustava *A* i *B*.

Provođenje testiranja može se odvijati na čitavoj populaciji, ili se može uzorkovati nad populacijom i ispitivati ponašanje sustava A i B na uzorcima x veličine n iz populacije X. U slučaju uzorkovanja, testiranje hipoteze procjenjuje kolika je vjerojatnost:

$$p(d_{A,B}(X) > d_{A,B}(x)|H_0) < \alpha \tag{4.2}$$

gdje je *X* slučajna varijabla mogućih dobivenih uzoraka veličine *n*, a $d_{A,B}(x)$ promatrana (konstantna) vrijednost. Ako vrijedi 4.2 za odabranu vrijednost α (najčešće 0.05) onda se odbacuje nulta hipoteza. Dva su osnovna načina izračuna p-vrijednosti:

– izravnim izračunom,

– procjenom.

U idućem poglavlju bit će govora o oba načina. Prvo će se pokazati na primjerima često vidljivim u časopisima o strojnom učenju kako je moguće izračunati p-vrijednost s obzirom na različite okolnosti eksperimenta. Na čitatelju ostaje prepoznati uvjete vlastitog istraživanja te se prema tome odlučiti za jedan od ova dva načina.

5. Izračun p-vrijednosti

P-vrijednost može se izračunati ili, ukoliko se ne može izračunati, aproksimirati. Jedna od najčešće korištenih metoda za procjenjivanje p-vrijednosti je upareni (engl. *bootstrap*). Upareni bootstrap jedna je od najčešće korištenih metoda (Koehn, 2004) zato što se može primjeniti na sve mjerne metode (uključujući složenije kao što su BLEU (Papineni et al., 2002), F1).

Konstruirano je M klasifikatora koji rade nad N različitih skupova podataka. Cilj je napraviti usporedbe:

- dvaju klasifikatora A i B na istom skupu podataka X,
- dvaju klasifikatora A i B na N različitih skupova podataka ,
- više od dva (M) klasifikatora na istom skupu podataka X te
- više od dva (M) klasifikatora na N različitih skupova podataka.

Prvo će se objasniti metode izravnog izračuna p-vrijednosti, a zatim metode kojima se aproksimira p-vrijednost.

5.1. Izravan izračun p-vrijednosti

5.1.1. Usporedba dvaju klasifikatora na istom skupu podataka

U radu (Dietterich, 1998) uspoređuje se pet statističkih testova usporedbe rezultata klasifikacije. Rezultati testova eksperimentalno se

5.1.2. Usporedba dvaju klasifikatora na različitom skup podataka

Analizom objavljenih radova na konferencijama *International Conference of Machine Learning* između 1999. i 2003. godine (Demšar, 2006) detektirane su metode usporedbi klasifikatora nad različitim skupovima podataka:

- prosjek preciznosti u različitim skupovima podataka,

– t-test,

– upareni t-test jedan protiv ostalih,

– upareni t-test svatko protiv svakog,

– prebrojavanje pobjeda/nerješenih rezultata/poraza i

– prebrojavanje statistički značajnih pobjeda/nerješenih rezultata/poraza

Opće je prihvaćeno da je korištenje t-testa za usporedbu klasifikatora na različitim skupovima podataka neprimjeren. Samo neki od značajnih radova koji odbacuju t-test zbog neprimjerenosti koncepta. U radu (Demšar, 2006) preporučuje se Wilcoxonov test rangova s predznacima (engl. *Wilcoxon ranked sign test*) (Wilcoxon, 1945). Uz njega, koristi se i test s predznacima Dixon i Mood (1946).

U idućim pododjeljcima opisat će se različiti načini mjerenja statističke značajnosti. Za svaki bit će kratki opis kako funkcionira, uz referencu na podrobniji opis, što se smatra prednostima i manama tog postupka u znanstvenoj zajednici.

Računanje prosjeka

Mjerenje performansi klasifikatora nad različitim skupovima podataka moguće je jednostavno zbrojiti i uzeti srednju vrijednost:

$$\hat{X} = \frac{1}{N} \sum_{i=1}^{N} c_i^j. \tag{5.1}$$

gdje je $\hat{X}$ je mjera dobrote klasifikacije, N broj skupova podataka, a c_i^j izvedba klasifikatora c^j (ovdje $j \in (1, 2)$) nad skupom podataka i.

(Webb, 2000) napominje kako ovakav način često nije smislen, jer usporedba grešaka između različitih skupova podataka iz različitih domena nije nešto što bi trebalo uspoređivati. Ukoliko nije smisleno raditi usporedbe rezultata u različitim domenama, zbranjanje tih vrijednosti također nije valjana operacija. Naravno, postoje slučajevi u kojima je ova metoda ispravna, stoga je ne valja u potpunosti odbaciti.

Upareni t-test

Upareni t-test (Sprinthall i Fisk, 1990) uzima u obzir prosječnu razliku u performansama nad različitim skupovima podataka i provjerava ima li statistički značajne razlike. Neka su c_i^1 i c_i^2 rezultati dva klasifikatora na skupu podataka i, od ukupno N skupova podataka. Razlika d_i se definira kao $c_i^1 - c_i^2$. Tada se *t* vrijednost se dobije $\hat{d}/\sigma_{\hat{d}}$, gdje

je $\sigma_{\hat{d}}$ varijanca varijable d_i raspoređenoj prema Studentovoj raspodjeli (engl. *Student distribution*) s $N-1$ stupnjeva slobode.

Tri su temeljne zamjerke korištenja t-testa pri evaluaciji klasifikatora. Prva zamjerka spomenuta je prilikom mjerenja značajnosti računanjem prosjeka. U oba slučaja rade se usporedbe performansi između različitih skupova podataka, ali u ovom slučaju uzima se u obzir varijanca između skupova podataka, što donekle umanjuje standardnu pogrešku σ_d, no i dalje se radi dvojbena usporedba različitih skupova podataka. U slučaju da je uzorak manji od (približno) 30 jedinki, korištenje t-testa zahtjeva da uzorak zadovoljava uvjete normalne razdiobe. Dakle, problematično je korištenje t-testa sa malim uzorcima, jer nije moguće pouzdano ispitati normalnost nad malim skupom podataka (Razali i Wah, 2011). Treći problem korištenja t-testa je problem ekstremnih vrijednosti. Ekstremne vrijednosti (jako dobri ili jako loši rezultati) mogu učiniti pomaknuti distribuciju rezultata i tako umanjiti snagu t-testa.

Wilcoxonov test rangova s predznacima

Wilcoxonov test rangova s predznacima (Wilcoxon, 1945) je neparametarska verzija uparenog t-testa. Razlike u performansama klasifikatora se rangiraju za svaki skup podataka. Rangovi se formiraju prema apsolutnim vrijednostima razlika, a potom se uspoređuju rangovi pozitivnih i negativnih razlika. U slučaju jednakih vrijednosti, u rang se upisuje prosječna vrijednost. Kao što smo opisali kod uparenog t-testa, d_i će predstavljati razliku performansi dva klasifikatora na i-tom skupu podataka (od ukupno N). Ako razlikujemo klasifikatore 1 i 2, potrebno je zbrojiti rangove R^1 i R^2, prema formulama:

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

$$R^2 = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

.

Nerješeni slučajevi se jednoliko raspoređuju u svaku sumu. Ukoliko postoji neparan broj nerješenih slučajeva, jedan nerješeni slučaj se zanemaruje, kako bi brojka uvijek bila jednaka. Rezultat je statistički značajan ukoliko je je dobivena vrijednost veća od granične T vrijednosti. T vrijednost se dobiva kao $min(R^1, R^2)$ Za uzorke manje od 25 moguće je pronaći točnih graničnih T vrijednosti (Wilcoxon et al., 1973), a za

broj uzoraka	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$\alpha = 0.10$	0	0	0	1	1	1	2	2	3	3	3	4	4	5	5	5	6	6	7	7	7

Tablica 5.1: Kritične vrijednosti za dvostrani test znakova. Ukoliko je broj pobjeda manji ili jednak onome iz drugog retka, razlika je statistički značajna.

ostale potrebno je izračunati spomenutu z statistiku:

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

Wilcoxonov test također radi usporedbe različitih skupova podataka, ali ne količinski, zbrajanjem iznosa razlika, već kvalitativno, zanemarujući apsolutne iznose razlika. Pretpostavka da su podatci normalno distribuirani nije potrebna, jer je Wilcoxonov test neparametarski. Utjecaj graničnih vrijednosti je smanjen jer se sada promatraju rangovi, umjesto apsolutnih vrijednosti.

Wilcoxonov test ima manju snagu od uparenog t-testa kada su podatci normalno distribuirani. U suprotnome, Wilcoxonov test može, ali i ne mora prema (Demšar, 2006) biti jači od t-testa.

Test znakova

Zanemarivanjem vrijednosti razlika u performansama klasifikatora isptivanje značajnosti svodi se na bilježenje rezultata kroz pobjede, poraze i izjednačene rezultate. Ako se uspoređuju dva algoritma na ovaj način, nulta hipoteza pretpostavlja kako će svaki algoritam imati približno $N/2$ pobjeda u skupu od N podataka. Pošto razlikujemo dva ishoda, podatke moguće je pretpostaviti binomnu distribuciju podataka (Miller et al., 1965) i provesti binomni test (Dixon i Mood, 1946), test znakova (engl. *sign test*). Postoje izračunate vrijednosti za test znakova vidljive u literaturi (Wilcoxon et al., 1973) prikazane i objašenjene u tablici 5.1.

U testu znakova ne rade se usporedbe između različitih skupova podataka, dakle nije potrebno zadovoljiti pretpostavku da se podatci mogu/smiju uspoređivati, niti je potrebno da podliježu normalnoj razdiobi. Mana testa znakova jest da neće odbaciti nultu hipotezu ukoliko rezultati jednog algoritma nisu konstatno bolji od drugog. Prema tome, snaga testa znakova je manja od Wilcoxonovog testa.

Prilikom nabrajanja korištenih testova statističke značajnosti u pododjeljku 5.1.2 još je spomenuto i prebrojavanje statistički značajnih pobjeda/nerješenih rezultata/poraza, što je ekvivalentno ovom načinu, prethodno filtrirajući rezultate nad kojima je utvrđena

statistička značajnost. Statistička značajnost ispituje se jednim od testova značajnosti (opisanim u poglavlju). Prema (Demšar, 2006) ovakav postupak je neispravan, jer pretpostavlja kako je moguće da statistički testovi razlikuju prave od slučajnih razlika. Statistički testovi mjere vjerojatnost dobivenog rezultata pod uvjetom da je nulta hipoteza zadovoljena, što nije ekvivalentno vjerojatnosti nulte hipoteze.

5.1.3. Usporedba više klasifikatora na različitom skupu podataka

Usporedba više od dva klasifikatora podrazumijeva kako više nije moguće provoditi testove zasnovane na uparenom t-testu iz istog razloga zašto nije moguće provoditi uzastopne t-testove – gomilanje pogreške prvog tipa. (Salzberg, 1997) je primjetio i dokazao kako je moguće primjeniti *Bonferonnijevu metodu* (engl. *Bonferroni correction*) i tako omogućiti višestruko testiranje. Bonferronijeva korekcija dozvoljava višestruke usporedbe kontroliranjem razine pogreške. (Bland i Altman, 1995). Primjenom korektivnog faktora Bonferonni korekcija umanjuje p smanjujući mogućnost pogreške tipa I.

Opće statističke metode koje bi se mogle koristiti za usporedbu više klasifikatora na različitim skupovima su analiza varijance (engl. *analysis of variance*) ANOVA i Friedmanov test (engl. *Friedman test*).

Analiza varijance

ANOVA je često korištena metoda za testiranje značajnosti razlika između više od dvije srednje vrijednosti. Opisana je u (Fisher, 1956). Nulta hipoteza koju ANOVA pretpostavlja je da nema statistički značajne razlike između vrijednosti koje se uspoređuju. Odbacivanje nulte hipoteze nam govori samo da postoji značajna razlika, a ne gdje se nalazi. U tu svrhu osmišljeni su različiti *post-hoc* testovi, primjerice Tukeyev (engl. *Tukey's test*) ili Dunnetov test (engl. *Dunnett test*)(Wallenstein et al., 1980). Tukeyev test uparuje i uspoređuje sve kombinacije klasifikatora, dok Dunnetov test radi usporedbe svih klasifikatora s jednim referentim. Navedeni *post-hoc* testovi u suštini se ne razlikuju od višestruke primjene t-testa, s time što *post-hoc* testovi kontroliraju mogućnost pogreške tipa I.

Korištenje ANOVA postupka zahtjeva dva uvjeta. ANOVA pretpostavlja kako analizirani uzorci podliježu normalnoj raspodjeli, što često nije slučaj prilikom analize postupaka strojnog učenja. Druga pretpostavka ANOVE je homogenost varijanci (engl. *homogeneity of variance*). Svojstvo homogenosti varijanci je ispunjeno ukoliko sve mjere imaju jednake varijance, što se često ne može primjeniti na rezultate klasi-

fikatora. Primjena *post-hoc* testova se ne preporučuje ukoliko pretpostavke ANOVE nisu zadovoljene (Zar, 1974).

Friedmanov test

Kao što je Wilcoxonov test znakova neparametarska verzija uparenog t-testa, tako je Friedmanov test (Friedman, 1937) neparametarska verzija višestruke ANOVE. U Friedmanovom testu se rangiraju algoritmi za svaki skup podataka zasebno. U slučaju jednakih vrijednosti, dodjeljuju se prosječni rangovi.

Ako je r_i^j rang j-tog klasifikatora na i-tom skupu podataka (ukupno N). Friedmanov test uspoređuje prosječne rangove klasifikatora:

$$R_j = \frac{1}{N} \sum_i^N r_i^j$$

Nulta hipoteza u Friedmanovom testu tvrdi da nema razlike između srednjih vrijednosti rangova klasifikatora. Iz ranga R_j, stupnjeva slobode $k-1$ Friedmanova vrijednost računa se:

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

.

U praksi se češće Friedmanova vrijednost svodi na F-vrijednost koja podliježe F-distribuciji s $k-1$ i $(k-1)(N-1)$ stupnjeva slobode prema formuli:

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

Friedman je usporedio svoj test s ANOVA-om kroz eksperiment (Friedman, 1940) i zaključio kako se testovi uglavnom slažu oko rezultata.

Kada Friedmanov test opovrgne nultu hipotezu, potrebno je nastaviti s *post-hoc testovima*. Nemenyijev test (engl. *Nemenyi test*) (Nemenyi, 1963) neparametarska je verzija Tukeyevog testa za ANOVA-u, dakle uspoređuje rezultate svih klasifikatora međusobno. Performanse dvaju klasfikatora se značajno razlikuju ukoliko se prosjeci njihovih rangova razlikuju za barem:

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

U slučaju da je potrebno uspoređivati sve klasifikatore s jednim referentnim, moguće je koristiti proceduru koja primjenom Bonferonnijevog korektivnog faktora radi međusobne usporedbe. U ovom slučaju radi se manje usporedbi $(k-1)$, naspram $k(k-1)/2$

prilikom izvedbe Nemenyjevog testa. Što se radi manje usporedbi, to je test jači, zato se i prilikom usporedbi s referentnim klasifikatorom preferira ovakav način usporedbe. Usporedba i-tog i j-tog klasifikatora se računa:

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

. Dobivena z vrijednost se uspoređuje s odgovarajućim vrijednostima normalne distribucije za utvrđivanje značajnosti.

U radu (Demšar, 2006) analizirani su Nemenyijev test s testom koji koristi Bonferonni korekciju. Zaključeno je kako je snaga testa veća ukoliko se uspoređuju svi klasifikatori s jednim referentnim. Međusobna usporedba svih klasifikatora se jedino preporučuje u slučaju kada je potrebno dokazati da su novootkrivene tehnike bolje od svih postojećih.

6. Zaključak

Zaključak.

7. Literatura

Taylor Berg-Kirkpatrick, David Burkett, i Dan Klein. An empirical investigation of statistical significance in nlp. U *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, stranice 995–1005. Association for Computational Linguistics, 2012.

Maximilian Bisani i Hermann Ney. Bootstrap estimates for confidence intervals in asr performance evaluation. U *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, svezak 1, stranice I–409. IEEE, 2004.

J Martin Bland i Douglas G Altman. Multiple significance tests: the bonferroni method. *BMJ: British Medical Journal*, 310(6973):170, 1995.

Nancy Chinchor. The statistical significance of the muc-4 results. U *Proceedings of the 4th conference on Message understanding*, stranice 30–50. Association for Computational Linguistics, 1992.

Michael Collins. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637, 2003.

Reuven Dar, Ronald C Serlin, i Haim Omer. Misuse of statistical tests in three decades of psychotherapy research. *Journal of consulting and clinical psychology*, 62(1):75, 1994.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Wilfrid J Dixon i Alexander M Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

Ronald A Fisher. Statistical methods and scientific inference. 1956.

Ronald Aylmer Fisher. The design of experiments. 1935.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200): 675–701, 1937.

Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

Mark A Hall. *Correlation-based feature selection for machine learning*. Doktorska disertacija, The University of Waikato, 1999.

Larry V Hedges, Ingram Olkin, Mathematischer Statistiker, Ingram Olkin, i Ingram Olkin. Statistical methods for meta-analysis, 1985.

Darrell Huff. *How to lie with statistics*. WW Norton & Company, 2010.

Philipp Koehn. Statistical significance tests for machine translation evaluation. U *EMNLP*, stranice 388–395, 2004.

Irwin Miller, John E Freund, i Richard Arnold Johnson. *Probability and statistics for engineers*, svezak 4. Prentice-Hall Englewood Cliffs, NJ, 1965.

Peter Nemenyi. *Distribution-free multiple comparisons*. Doktorska disertacija, Princeton University, 1963.

Kishore Papineni, Salim Roukos, Todd Ward, i Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. U *Proceedings of the 40th annual meeting on association for computational linguistics*, stranice 311–318. Association for Computational Linguistics, 2002.

John D Potter. The lady tasting tea: How statistics revolutionized science in the twentieth century. *Nature Medicine*, 7(8):885–886, 2001.

DMW Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Mary Louise Pratt. *Toward a speech act theory of literary discourse*, svezak 266. Indiana University Press Bloomington, 1977.

Nornadiah Mohd Razali i Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

Steven L Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and knowledge discovery*, 1(3):317–328, 1997.

Richard C Sprinthall i Stephen T Fisk. *Basic statistical analysis*. Prentice Hall Englewood Cliffs, NJ, 1990.

Bruce Thompson. The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Educational*, stranice 361–377, 1993.

SYLVAN Wallenstein, Christine L Zucker, i JOSEPH L Fleiss. Some statistical methods useful in circulation research. *Circulation Research*, 47(1):1–9, 1980.

Geoffrey I Webb. Multiboosting: A technique for combining boosting and wagging. *Machine learning*, 40(2):159–196, 2000.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1 (6):80–83, 1945.

Frank Wilcoxon, SK Katti, i RA Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1973.

Alexander Yeh. More accurate tests for the statistical significance of result differences. U *Proceedings of the 18th conference on Computational linguistics-Volume 2*, stranice 947–953. Association for Computational Linguistics, 2000.

Nigel G Yoccoz. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, 72(2):106–111, 1991.

Jerrold H Zar. Multiple comparisons. *Biostatistical analysis*, 1:185–205, 1974.

Ying Zhang, Stephan Vogel, i Alex Waibel. Interpreting bleu/nist scores: How much improvement do we need to have a better system? U *LREC*, 2004.