*Elektrotehniški vestnik 73(5): 279–284, 2006*
*Electrotechnical Review, Ljubljana, Slovenija*

# Statistične primerjave klasifikatorjev pri strojnem učenju

**Janez Demšar**

*Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana, Slovenija*
*E-pošta: janez.demsar@fri.uni-lj.si*

**Povzetek.** Statistična primerjava klasifikatorjev je eden ključnih elementov pri raziskovanju novih metod strojnega učenja ali izboljšav obstoječih. Kljub pomembnosti takšnih primerjav pa to področje še ni dobro raziskano, zato se v praksi pogosto uporabljajo testi, katerih primernost je v najboljšem primeru dvomljiva. Prispevek vsebuje pregled statističnih testov, ki se uporabljajo ali pa bi se jih dalo uporabiti za primerjanje klasifikatorjev. Pri tem smo opazovali predvsem njihovo primernost za naše potrebe z vidika tega, kaj merijo in kakšne so njihove predpostavke o podatkih. Osrednji del članka je empirična primerjava testov na izboru nekaj popularnih metod učenja in pogosto uporabljanih testnih podatkov. Tako teoretični kot empirični del študije kažeta, da so za primerjanje klasifikatorjev najprimernejši neparametrični testi, saj ne temeljijo na (verjetno kršeni) predpostavki normalnosti, ne zahtevajo primerljivosti uspešnosti klasifikatorjev na različnih domenah, poleg tega pa so, kot je videti, močnejši od parametričnih.

**Ključne besede:** strojno učenje, umetna inteligenca, statistični testi

# Statistical Comparisons of Classifiers in Machine Learning

**Extended abstract.** Comparisons between classifiers are a crucial element in most studies that introduce new machine learning algorithms or modifications of the existing ones. Despite their importance, there is no consensus in the community regarding which test should be applied in a certain situation, and excellent machine learning papers quite often conclude with statistical tests that are conceptually or statistically inappropriate. The situation is especially bad in comparisons of multiple classifiers, where the tests designed for comparisons of two samples are often used on each pair of classifiers instead of using omnibus tests like ANOVA or at least applying the appropriate corrections for multiple hypotheses testing.

We analyzed the tests which are or which should (in our opinion) be used in machine learning studies: the paired t-test, the Wilcoxon signed-ranks test and the sign test for comparison of two classifiers, and ANOVA and the Friedman test with appropriate post-hoc tests for comparisons of multiple classifiers [10]. We checked what the tests really measure and what assumptions they make about the data; specifically, the parametric tests require commensurability of the results accross different domains and assume that the results of classifiers are distributed normally. Since both of these conditions for the use of parametric tests are most probably violated, we dissuade from using the parametric tests. On the other hand, the described non-parametric tests suffer from none of these deficiencies.

The same conclusion in favour of non-parametric tests is reached in the experimental part of the paper where we compare the tests on a selection of standard machine learning algorithms using the data sets from the UCI machine learning repository [1]. The non-parametric tests seem to have more power and be more replicable than the parametric ones both in comparisons between two (Fig. 1) and between multiple classifiers (Fig. 3), with the only exception of the Dunnet test for post-hoc comparisons of one classifier against all others, which rejects a somewhat larger number of hypotheses than the corresponding non-parametric alternative.

Altogether, we recommend the use of non-parametric tests for comparisons of classifiers, but warn that other criteria beyond the grasp of statistics should be considered and possibly even favoured over the pure improvements in predictive power of classifiers.

**Key words:** machine learning, artificial intelligence, statistical tests

*Prejet 20. februar, 2006*
*Odobren 5. december, 2006*

## 1 Uvod

Večina študij novih ali izboljšanih metod strojnega učenja vsebuje tudi primerjavo teh metod z obstoječimi metodami. Kljub pomembnosti takšnih primerjav pa je bilo doslej temeljito obdelano zgolj vprašanje primerjav dveh klasifikatorjev na eni sami problemski domeni [9], ne pa tudi veliko pogostejša situacija, v kateri dva ali več klasifikatorjev primerjamo na večji množici, tipično desetih do dvajsetih domen. Demšar [3] je v analizi prispevkov, sprejetih na srečanjih International Conference of Machine Learning med leti 1999 in 2003, pokazal, da glede teh testov ni ustaljene prakse, temveč se avtorji odločajo za različne metode in teste značilnosti, med katerimi pa so številni z vidika statistične korektnosti vsaj neprimerni, če ne kar napačni.

Problem, ki ga obravnavamo v članku, je naslednji. Denimo, da smo za potrebe primerjave testirali algoritme učenja na več problemskih domenah (zbirkah podatkov, različnih učnih problemih). Naj bo $c_i^j$ rezultat $j$-tega izmed $k$-tih algoritmov na $i$-ti izmed $n$-tih domen. Ničelna hipoteza, ki jo želimo na podlagi teh meritev preveriti in po možnosti zavrniti, je, da so vsi klasifikatorji enako uspešni. Pri primerjavah več kot dveh klasifikatorjev želimo izvedeti tudi, kateri izmed njih se značilno razlikujejo.

Uporabljena mera uspešnosti klasifikatorjev in način vzorčenja podatkov za razpravo nista pomembna. Zahtevamo le, da so izmerjeni rezultati zanesljivi, kar v praksi pomeni, da morajo biti domene primerno izbrane in da jih mora biti dovolj.

Obravnavane metode bodo uporabljale zgolj izmerjene rezultate $c_i^j$, ne pa tudi njihove variance $\sigma_{c_i^j}$ znotraj posamezne domene pri posameznem klasifikatorju. Zato se nam pri analizi ne bo treba ubadati s podcenjevanjem variance, do katere pride pri prečnem preverjanju in sorodnih metodah vzorčenja ter je glavna težava pri primerjavah klasifikatorjev na eni sami domeni.

V članku bomo najprej analizirali nekaj statističnih metod za primerjanje klasifikatorjev z vidika njihovih predpostavk in statističnih lastnosti. Nato bomo praktično ocenili njihovo moč in ponovljivost tako, da bomo opazovali njihove rezultate na izboru nekaj popularnih metod strojnega učenja in množici podatkov z repozitorija na UCI [1]. Na podlagi teoretičnih in empirično opaženih lastnosti testov bomo v sklepu priporočili najprimernejše teste za primerjanje dveh ali več algoritmov učenja.

## 2 Teoretična obravnava testov

V tem razdelku bomo predstavili nekaj testov, ki se ali pa bi se lahko uporabljali za primerjavo klasifikatorjev. V obravnavi bomo ločeno opazovali teste za primerjavo dveh in teste za primerjavo več klasifikatorjev.

### 2.1 Primerjave dveh klasifikatorjev

**Povprečenje prek domen.** V primerjavah klasifikatorjev nekateri avtorji računajo povprečno kakovost klasifikatorja prek vseh domen, $\overline{c^j} = \frac{1}{n}\sum_i c_i^j$. Kot ugovarja Webb [11], ocene na različnih domenah med seboj navadno niso primerljive, zato tovrstna povprečja nimajo smisla, prav tako pa tako izračunanih povprečij nima smisla uporabljati v statističnih testih, denimo, v t-testu. Takšno testiranje je na mestu zgolj, če gre za več zbirk podatkov iz sorodnih problemskih domen, denimo za analizo točnosti napovedovanja določene bolezni na podatkih, zbranih v različnih bolnišnicah.

**T-test po parih.** Za primerjanje dveh klasifikatorjev se pogosto uporablja t-test po parih [10], ki preverja, ali je povprečje razlik med klasifikatorjema, $\overline{d} = \frac{1}{n}\sum_i d_i$, kjer $d_i = c_i^1 - c_i^2$, značilno večje od variance teh razlik.

Tudi takšen t-test je po našem mnenju konceptualno neustrezen, ker še vedno temelji – čeprav navadno ni zapisan v takšni obliki – na povprečenju prek domen, saj je $\overline{d}$ enak $\overline{c^2} - \overline{c^1}$. Edina razlika med t-testom po parih in t-testom na navadnih povprečjih je v tem, da je prvi zaradi upoštevanja kovariance med klasifikatorjema močnejši v strogem smislu moči testov.

Webb [11] predlaga, da bi rešili problem neprimerljivosti tako, da bi namesto razlik med klasifikatorjema opazovali razmerja in namesto aritmetičnih računali geometrijska povprečja, $\prod_i c_i^1/c_i^2$. Ker je slednje enako $\exp\left(\sum_i(\ln c_i^1 - \ln c_i^2)\right)$, njegov predlog ne reši problema, saj le zamenja povprečje običajnih razlik s povprečjem razlik logaritmov. Drug, morda boljši način zagotavljanja primerljivosti, je opazovanje relativnih razlik, $d_i = (c_i^1 - c_i^2)/(c_i^1 + c_i^2)$.

T-test je neprimeren tudi z vidika predpostavk o podatkih, saj zahteva, da je opazovana spremenljivka porazdeljena normalno ali pa mora biti vzorec dovolj velik (običajna meja je vsaj 30). Število domen v primerjavah klasifikatorjev je tipično veliko manjše, obenem pa narava podatkov ne zagotavlja normalnosti. Majhno število podatkov obenem preprečuje preverjanje normalnosti s pomočjo testov, kot je test Kolmogorova-Smirnova.

T-test je občutljiv na izstopajoče primere: izjemno dobri rezultati algoritma na eni domeni uteg-

nejo prevladati nad slabimi rezultati na večini drugih in nasprotno, kar je v večini primerov nezaželeno.

**Wilcoxonov test predznačenih rangov** [12, 10] je neparametrična alternativa t-testu. Test rangira absolutne vrednosti razlik med uspešnostjo klasifikatorjev po absolutnih vrednostih in sešteje range, ki pripadajo domenam, pri katerih je bil boljši prvi ali drugi klasifikator. Po ničelni hipotezi bi morali biti vsoti enaki. Za ocenjevanje značilnosti odstopanja od enakosti lahko uporabljamo tabelirane kritične vrednosti za manjšo izmed vsot, pri velikem številu domen ($n > 25$) pa je le-ta porazdeljena približno po $N(n(n+1)/4, \sqrt{n(n+1)(2n+1)/24})$.

Wilcoxonov test nima nobene od naštetih slabosti t-testa. Ker temelji na rangiranju razlik, ne zahteva primerljivosti med uspešnostjo na posameznih domenah, obenem pa večje razlike vseeno štejejo več kot manjše. Kot neparametrični test ne predpostavlja normalne ali kake druge distribucije in ni občutljiv na izstopajoče primere. Kadar predpostavke, ki jih zahteva t-test, veljajo, je Wilcoxonov test šibkejši od t-testa, kadar so kršene, pa je Wilcoxonov test lahko celo močnejši.

**Test znakov** [12, 10] ali binomski test zavrne ničelno hipotezo o enakosti, če je število domen, na katerih je uspešnejši klasifikator boljši od manj uspešnega, značilno večje od $n/2$. Dejanske vrednosti razlik pri tem testu niso pomembne. Tako kot pri Wilcoxonovem testu pri manjšem številu domen uporabimo tabele kritičnih vrednosti, pri večjem pa normalno aproksimacijo in razglasimo razliko za značilno, če število zmag enega od klasifikatorjev presega $n/2 + 1.96\sqrt{n}/2$, kar je približno $n/2 + \sqrt{n}$.

Test znakov pogosto videvamo v primerjavah klasifikatorjev, čeprav je šibkejši od Wilcoxonovega testa in pred njim v našem primeru nima posebnih prednosti, razen te, da je zaradi preprostega izračuna primeren za hitro ročno presojo statistične značilnosti.

### 2.2 Primerjave več klasifikatorjev

Pri primerjavah med več klasifikatorji avtorji člankov pogosto testirajo razlike med vsakim parom klasifikatorjev. Z vidika statistike je to nedopustno, saj veliko število testiranih hipotez bistveno poveča možnost zavrnitve resnične ničelne hipoteze. Namesto tega statistika priporoča teste, razvite posebej za ta namen (ANOVA, Friedmanov test ipd.), ali pa zahteva, da pri testih po parih ustrezno zaostrimo kritične vrednosti z Bonferonijevim ali podobnim popravkom. V člankih s področja strojnega učenja le občasno videvamo slednje, ANOVA in posebej Friedmanov test pa sta uporabljena zelo redko.

**ANOVA** [5, 10] je znani statistični test, ki razdeli celotno varianco na varianco znotraj skupin in med njimi. V našem primeru je smiselno uporabiti ANOVO za povezane meritve, saj vse klasifikatorje vrednotimo na istih domenah. Podrobnejšo predstavitev postopka je mogoče najti v večini statističnih knjig.

Podobno kot t-test tudi ANOVA zahteva primerljivost meritev, normalne porazdelitve in, še pomembneje, sferičnost (kriterij podoben homogenosti variance pri običajni ANOVI) ter je občutljiva na izstopajoče primere. Argumenti proti uporabi ANOVE za primerjanje klasifikatorjev so torej enaki kot argumenti proti t-testu.

**Friedmanov test** [6, 10] je neparametrična zamenjava za zgornjo obliko ANOVE. Deluje tako, da rangira uspešnosti klasifikatorjev na vsaki domeni posebej in nato izračuna povprečni rang $R_i$ vsakega klasifikatorja prek vseh domen. Dobljene statistike so informativne same po sebi, saj povedo, neformalno, katero mesto med klasifikatorji v povprečju zasede posamezni klasifikator. Ničelna hipoteza predpostavlja, da so povprečni rangi $R_i$ enaki. Neenakost merimo s Friedmanovo statistiko $\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$, ki je pri dovolj velikih $N$ in $k$ (denimo $N > 10$, $k > 5$) porazdeljena približno po $\chi_{k-1}^2$; pri manjših pa so kritične vrednosti tabelirane. V praksi je primernejša manj konservativna Iman-Davenportova statistika $F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$, ki je porazdeljena po $F_{k-1,(N-1)(k-1)}$.

Razmerje med Friedmanovim testom in ANOVO je podobno kot med Wilcoxonovim in $t$-testom: Friedmanov test ne predpostavlja posebnih lastnosti porazdelitve in je, kadar te dejansko niso izpolnjene, lahko močnejši od ANOVE.

**Post-hoc testi.** ANOVA in Friedmanov test preverjata ničelno hipotezo, da so vsi klasifikatorji enako uspešni. Kadar jo zavrneta, praviloma uporabimo post-hoc test, s katerim določimo tiste pare klasifikatorjev, ki se razlikujejo med seboj. Pri ANOVI sta najprimernejša Tukeyev test za primerjanje vsakega klasifikatorja z vsakim in Dunnettov test za primerjanje enega z vsemi drugimi. Analogna testa po Friedmanovem testu sta Nemenyijev in Bonferroni-Dunnov test [10].

## 3 Empirična primerjava testov

Teste iz prejšnjega razdelka bi želeli preveriti s poskusom, v katerem bi vedeli, kdaj je ničelna hipoteza resnična in kdaj ne. Žal to ni praktično izvedljivo, saj bi za to potrebovali umetne podatke in klasifikatorje, pri sestavljanju katerih bi morali vedeti, na kakšen način in kako pogosto se pri klasifikaciji motijo. S

tem pa bi že določili tudi porazdelitev uspešnosti klasifikatorjev prek domen, torej bi nam o primernosti posameznih testov že vse povedala statistična teorija. Če je uspešnost porazdeljena, denimo, normalno, že vemo, da so za primerjave najprimernejši t-test, ANOVA in ustrezni post-hoc testi. V resnici pa želimo s testiranjem preveriti ravno, v kolikšni meri držijo predpostavke teh testov na resničnih podatkih in klasifikatorjih, oziroma kako kršitev le-teh vpliva na njihove rezultate.

Zato smo namesto umetnih klasifikatorjev primerjali štiri različne nastavitve C4.5 [8]: privzete vrednosti, nastavljanje minimalnega števila primerov v listih (`m`) z notranjim prečnim preverjanjem, nastavljanje stopnje zaupanja pri rezanju (`cf`) in nastavljanje obeh parametrov hkrati. Poleg tega smo preskušali še naivni Bayesov klasifikator s Fayyad-Iranijevo diskretizacijo zveznih atributov, naivni Bayesov klasifikator, ki verjetnosti pri zveznih atributih modelira z LOESS in metodo $k$ najbližjih sosedov ($k = 10$, sosedi so uteženi z Gausovskim jedrom). Klasifikatorje smo primerjali glede na njihovo klasifikacijsko točnost. Vsi poskusi so bili izvedeni s sistemom za strojno učenje Orange [4].

Klasifikatorje smo preskušali na 40 podatkovnih zbirkah* z repozitorija UCI [1]. V vsakem poskusu smo naključno izbrali po deset domen, pri čemer je bila verjetnost izbire $i$-te sorazmerna $(1 + e^{-kd_i})^{-1}$, kjer je $d_i$ vnaprej izmerjena (pozitivna ali negativna) razlika med uspešnostjo primerjanih klasifikatorjev na tej zbirki, $k$ pa parameter, s katerim lahko reguliramo razliko med klasifikatorjema oz. "napačnost" ničelne hipoteze. Čim večji je $k$, tem bolj so zbirke izbrane v korist enega izmed klasifikatorjev.† Takšno funkcijo smo izbrali zgolj zaradi prikladne oblike; dejanskega izbora ne moremo modelirati, poleg tega to za naš poskus ni posebej pomembno. V poskusih smo uporabljali vrednosti $k$ od 0 do 20 in z vsako po tisočkrat naključno izbrali po deset domen.

Ker v poskusu uporabljamo resnične klasifikatorje in domene, ne vemo, katere od preverjanih ničelnih hipotez bi morale biti zavrnjene. Zato smo namesto moči testov opazovali delež zavrnjenih hipotez pri stopnji značilnosti $\alpha = 0,05$ in ponovljivost, kakor jo je definiral Bouckaert [2]: pri vsakem $k$ smo prešteli, koliko parov izmed tisoč poskusov je neskladnih (pri eni izbiri domen je ničelna hipoteza zavrnjena in pri drugi ne). Meri sta povezani; če je zavrnjenih $q$ izmed $s$ testiranih hipotez, je ponovljivost enaka $[q(q-1) + (s-q)(s-q-1)]/[s(s-1)]$.

Obe meri sta odvisni od zahtevane stopnje značilnosti in, kar je še posebej moteče, pri merjenju ponovljivosti so bolje ocenjeni testi, pri katerih so pri danem $k$ stopnje značilnosti čim bolj oddaljene od tega praga. Zato smo poleg števila zavrnjenih hipotez in ponovljivosti po Bouckaertu opazovali tudi povprečno značilnost $p$, ki jo vrne test pri posameznem $k$, in njeno varianco. Slednjo smo odšteli od 1, tako da večji rezultat pomeni boljšo ponovljivost.

### 3.1 Primerjave dveh klasifikatorjev

Empirično smo primerjali štiri teste: t-test, t-test z relativnimi razlikami, Wilcoxonov test predznačenih rangov in test znakov. Slika 1 kaže rezultate primerjave med C4.5 z nastavljanjem `cf` in naivnega Bayesovega klasifikatorja z diskretizacijo. Pri vseh nastavitvah $k$ je največ hipotez zavrnil Wilcoxonov test, sledita mu t-test in t-test z relativnimi razlikami, najmanjkrat pa, skladno s pričakovanji, hipotezo zavrne test znakov. Enak je tudi vrstni red povprečnih značilnosti $p$.

Ponovljivost po Bouckaertu je tem manjša, čim bolj so $p$ vrednosti blizu izbranemu pragu. V tem pogledu je najbolj nezanesljiv Wilcoxonov test, vendar moramo to razumeti kot argument proti uporabljeni meri ponovljivosti in ne Wilcoxonovem testu. Opazovanje variance namreč pokaže, da $p$-vrednost najmanj niha ravno pri tem testu.

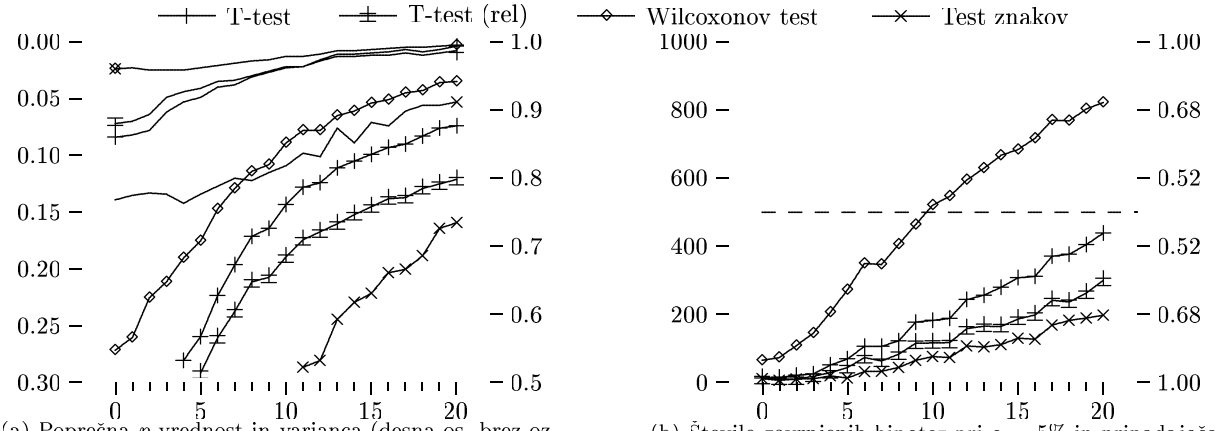
Na isti način smo primerjali tudi vse druge pare klasifikatorjev in vedno prišli do enakih rezultatov, z izjemo obeh $t$-testov, ki sta pogosto dajala tako rekoč enake rezultate prek vseh vrednosti $k$.

### 3.2 Primerjave več klasifikatorjev

Poskusi z ANOVO in Friedmanovim testom so dali podobne rezultate kot poskusi s testi za dva klasifikatorja: neparametrični Friedmanov test v povprečju vrača manjše $p$-vrednosti z manjšo varianco in večkrat zavrne ničelno hipotezo, vendar so razlike med testoma manjše kot pri primerjavah dveh klasifikatorjev (slika 2). Domene smo izbirali na podlagi razlik v klasifikacijski točnosti C4.5 in naivnega Bayesovega klasifikatorja, podobne rezultate pa smo dobili tudi z drugačno izbiro klasifikatorjev.

*adult, balance-scale, bands, breast cancer (haberman), breast cancer (lju), breast cancer (wisc), car evaluation, contraceptive method choice, credit screening, dermatology, ecoli, glass identification, hayes-roth, hepatitis, housing, imports-85, ionosphere, iris, liver disorders, lung cancer, lymphography, mushrooms, pima indians diabetes, post-operative, primary tumor, promoters, rheumatism, servo, shuttle landing, soybean, spambase, spect, spectf, teaching assistant evaluation, tic tac toe, titanic, voting, waveform, wine recognition, yeast

†V praksi raziskovalec navadno predpostavi, v katerih pogledih je njegova metoda boljša od metod, s katerimi jo primerja, zato bo izbral podatkovne zbirke, na katerih bo to razliko lahko pokazal, kriterij za izbiro pa v članku seveda obrazložil. Izbiranje zbirk glede na vnaprej izmerjeno točnost klasifikatorjev, tako kot to počnemo v tej raziskavi, bi bilo v resnični primerjavi klasifikatorjev nepošteno.




(a) Poprečna $p$-vrednost in varianca (desna os, brez oznak na krivuljah); Average $p$ values and their variance (right axis, no symbols on lines)

(b) Število zavrnjenih hipotez pri $\alpha = 5\%$ in pripadajoča ponovljivost (desna os); The number of rejected hypotheses at $\alpha = 5\%$ and the related consistency (right axis)

Slika 1. Rezultati testov za primerjanje dveh klasifikatorjev; Test results allowing for a comparison of two classifiers

Tudi med post-hoc testi za medsebojno primerjavo vseh klasifikatorjev (slika 3) je neparametrični Nemenyijev test videti močnejši od parametričnega Tukeyevega. Pri primerjavi enega klasifikatorja z ostalimi je Dunnetov test prehitel Bonferonijevega, kar lahko razložimo s tem, da je Dunnetov test sestavljen posebej za tovrstne post-hoc primerjave, Bonferonijev test pa temelji na splošni korekciji za testiranje več hipotez in je znan kot pretirano konservativen.

## 4 Sklep

Tako teoretična analiza testov kot praktični poskusi kažejo, da so za primerjanje klasifikatorjev v strojnem učenju neparametrični testi primernejši od parametričnih. Parametrični testi predpostavljajo primerljivost rezultatov na različnih domenah, zahtevajo normalno porazdelitev le-teh in so občutljivi na izstopajoče primere. Neparametrični testi teh pomanjkljivosti nimajo. Čeprav so neparametrični testi načelno šibkejši od parametričnih, naši poskusi kažejo drugače, kar namiguje, da so predpostavke parametričnih testov na teh problemih dejansko kršene.

Korektna statistična analiza poveča težo rezultatov primerjav klasifikatorjev, vendar zgolj zavrnitev ali nezavrnitev ničelne hipoteze ne sme biti ključni dejavnik za odločanje o uspešnosti novih metod v strojnem učenju. Uporabnost v praksi, preprostost vizualizacije, razložljivost modelov in celo subjektivna privlačnost metod ter podobni kriteriji zunaj dosega statističnih testov si gotovo zaslužijo večjo težo kot golo izboljšanje napovedne točnosti.

## 5 Literatura

[1] C.L. Blake, C.J. Merz, UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, 1998.

[2] R. R. Bouckaert, Estimating Replicability of Classifier Learning Experiments, V C. Brodley, *Machine Learning, Proceedings of the Twenty-First International Conference (ICML 2004)*, Menlo Part, CA, ZDA, AAAI Press, 2004.

[3] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, *Journal of Machine Learning Research 7*, 2006, 1–31.

[4] J. Demšar, B. Zupan, *Orange: From Experimental Machine Learning to Interactive Data Mining, A White Paper*, Faculty of Computer and Information Science, Ljubljana, Slovenia, 2004.

[5] R. A. Fisher, *Statistical methods and scientific inference (2nd edition)*, New York, Hafner Publishing Co., 1959.

[6] 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)*, 675–701, 1937.

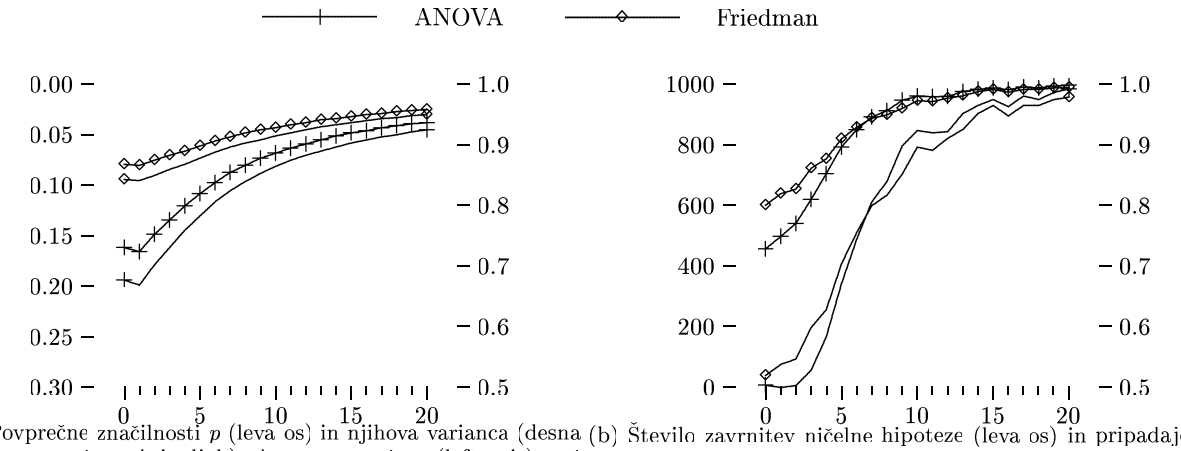
[7] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology 143(1)*, 29–36, 1982.

[8] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Francisco, Morgan Kaufmann Publishers, 1993.

[9] S. L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery (1)*, 317–328, 1997.

[10] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman & Hall/CRC, 2000.

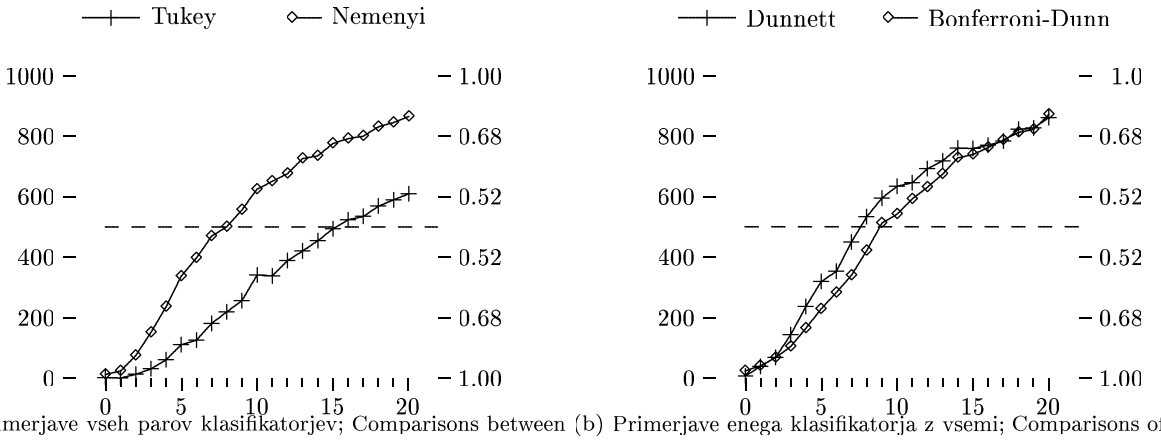
[11] G. I. Webb, MultiBoosting: A Technique for Combining Boosting and Wagging, *Machine Learning 40*, 2000, 159–197.




(a) Povprečne značilnosti $p$ (leva os) in njihova varianca (desna os, brez oznak na krivuljah); Average $p$ values (left axis) and their variance (no symbols on lines, right axis)

(b) Število zavrnitev ničelne hipoteze (leva os) in pripadajoča ponovljivost (desna os, brez oznak); Number of experiments in which the null-hypothesis was rejected (left axis) and the corresponding replicability (no symbols on lines, right axis)

Slika 2. Primerjava ANOVE in Friedmanovega testa; Comparison between ANOVA and the Friedman test




(a) Primerjave vseh parov klasifikatorjev; Comparisons between all pairs of classifiers

(b) Primerjave enega klasifikatorja z vsemi; Comparisons of one classifier with all others

Slika 3. Primerjava post-hoc testov; Comparison of the post-hoc tests

[12] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics 1*, 80–83, 1945.