# *Ocenjevanje zanesljivosti posameznih napovedi pri nadzorovanem učenju*

Darko Pevec

doktorska disertacija

predana

Fakulteti za računalništvo in informatiko

kot del izpolnjevanja pogojev za pridobitev naziva

doktor znanosti

s področja

računalnišva in informatike

Ljubljana, 2013

# IZJAVA O AVTORSTVU
## DOKTORSKE DISERTACIJE

*Izjavljam, da sem avtor doktorske disertacije z naslovom*

*Ocenjevanje zanesljivosti posameznih napovedi pri nadzorovanem učenju,*

*ki sem jo izdelal samostojno pod vodstvom mentorja in da slednje ne vsebuje materiala, ki bi ga kdorkoli predhodno že objavil ali oddal v obravnavo za pridobitev naziva na univerzi ali na drugem visokošolskem zavodu, razen v primerih, kjer so navedeni viri. Elektronska oblika doktorske disertacije je identična s tiskano obliko doktorske disertacije. Soglašam z javno objavo elektronske doktorske disertacije.*

— Darko Pevec —
junij 2013

Oddajo so odobrili

dr. Igor Kononenko
*redni profesor za računalništvo in informatiko*
MENTOR IN ČLAN OCENJEVALNE KOMISIJE

dr. Marko Robnik-Šikonja
*izredni profesor za računalništvo in informatiko*
PREDSEDNIK OCENJEVALNE KOMISIJE

dr. Nada Lavrač
*redna profesorica za računalništvo*
ZUNANJI ČLAN OCENJEVALNE KOMISIJE
Institut Jožef Stefan

# PREDHODNA OBJAVA

Izjavljam, da so bili rezultati obravnavane raziskave predhodno objavljeni/sprejeti za objavo v recenzirani reviji ali javno predstavljeni v naslednjih primerih:

[1] D. Pevec in Z. Bosnić. *Estimating reliability of single classifications*. Vabljeno predavanje, Laboratório de Inteligência Artificial e Apoio à Decisão, Universidade do Porto, Porto, 24/6/2010.

[2] D. Pevec. *Input dependent prediction intervals for arbitrary regression models*. Vabljeno predavanje, Laboratório de Inteligência Artificial e Apoio à Decisão, Universidade do Porto, Porto, 15/11/2011.

[3] D. Pevec in I. Kononenko. Model Selection with Combining Valid and Optimal Prediction Intervals. V zborniku *2012 IEEE 12th International Conference on Data Mining Workshops*, str. 653–658, 2012.

[4] D. Pevec in I. Kononenko. Input dependent prediction intervals for supervised regression. *Intelligent Data Analysis*, 18(5), 2014, v tisku.

Potrjujem, da sem pridobil pisna dovoljenja vseh lastnikov avtorskih pravic, ki mi dovoljujejo vključitev zgoraj navedenega materiala v pričujočo disertacijo. Potrjujem, da zgoraj navedeni material opisuje rezultate raziskav, izvedenih v času mojega podiplomskega študija na Univerzi v Ljubljani.

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*
Darko Pevec
*Ocenjevanje zanesljivosti posameznih napovedi pri nadzorovanem učenju*

# POVZETEK

Disertacija obravnava ocenjevanje zanesljivosti posameznih napovedi pri nadzorovanem učenju. Gre za relativno mlado področje strojnega učenja, trenutno pa mu strokovnjaki in uporabniki ne namenjajo dovolj pozornosti. Glavni kazalci uspešnosti modelov strojnega učenja so namreč še vedno mere povprečnega delovanja, na primer povprečna klasifikacijska točnost in koren srednje kvadratične napake. Vendar povprečne mere še zdaleč ne nudijo popolne slike in za vse več današnjih uporabnikov metod strojnega učenja so poleg samih napovedi zanimive tudi druge informacije, ki lahko pripomorejo k boljšemu razumevanju delovanja modelov. Dodatne informacije so zlasti pomembne, kadar imajo napačne napovedi lahko hude ekonomske ali zdravstvene posledice. Takrat strokovnjaki napovednim sistemom ne zaupajo zlahka, če jim le-ti ne nudijo dodatnih informacij o zanesljivosti. Ustaljene metode ocenjevanja točnosti napovednih sistemov so nezadostne in na področjih, kjer je podpora pri odločanju pomembna oziroma kjer povprečna uspešnost ni najpomembnejša, so informacije o zanesljivosti posameznih napovedi lahko zelo koristne.

Pri uporabi metod nadzorovanega strojnega učenja so vprašanja, kot so ali ima izbrani model dobro prestavitev podatkov in ali so napovedane vrednosti konformne zelo pomembna, saj se modeli zlahka naučijo napačnih konceptov ali pa se preveč prilagodijo šumu v podatkih. Ker želimo zajeti vse modele strojnega učenja, moramo nanje gledati iz zelo splošnega vidika – kot na črne škatle, za katere poznamo le njihov vhod (učne podatke) in njihov izhod ob poljubnih podatkih (testni podatki, novi primeri).

Delo nudi zaokrožen pregled področja ocenjevanja zanesljivosti pri nadzorovanem učenju. Nadzorovano učenje se zaradi inherentnih razlik med diskretnim in zveznim deli na klasifikacijo in regresijo. Cenilke zanesljivosti posameznih napovedi lahko ločimo na točkovne in intervalne. Zanimivo je, da so točkovne cenilke uporabne tako pri klasifikaciji kot pri regresiji, medtem ko so intervalne cenilke definirane le na regresij-

skih problemih.

Prvi pomemben prispevek disertacije je nova primerjalna študija uporabnosti točkovnih cenilk na klasifikacijskih primerih. Analiza z uporabo referenčne ocene kaže na to, da so točkovne ocene zanesljivosti koristne le v redkih primerih, na primer takrat, ko smo omejeni na uporabo ne optimalnih modelov, točkovne ocene zanesljivosti pa se dobro prilegajo podatkom zadanega problema.

V tem delu smo prevedli obstoječe pristope intervalnega ocenjevanja zanesljivosti posameznih napovedi na skupni imenovalec in s tem omogočili njihovo primerjavo, ki prej ni bila mogoča. Analiza razkriva dualno naravo intervalnih cenilk, v smislu da družina metod na osnovi stremljenja in maksimalnega verjetja optimizira pravilnost, druga družina metod na osnovi lokalnih okolic pa optimalnost. V delu predstavljamo kombinirano metodo, ki združuje lastnosti obeh družin. Naša testiranja kažejo na to, da kombinirani pristop nudi bolj robustne napovedne intervale.

Obstoječe statistike, ki nudijo informacije o povprečni uspešnosti modelov, niso dovolj informativne, po drugi strani pa so ustrezni grafični prikazi zelo koristni za razvoj uporabnikove intuicije in razumevanja delovanja modelov. Zato smo, poleg vizualizacijske tehnike za primerjavo napovednih intervalov več pristopov, predstavili novo vizualizacijsko tehniko, ki omogoča odkrivanje novih zakonitosti v podatkih ter vizualno primerjavo in evalvacijo več modelov.

Kot zadnji prispevek smo na osnovi statistik uspešnosti intervalnih ocen predlagali postopek oziroma novo kombinirano statistiko, ki omogoča robustno izbiro in združevanje regresijskih napovedi. Novi način ocenjevanja modelov in združevanja napovedi nudi prilegajoče in posledično bolj zanesljive napovedi.

*Ključne besede:* nadzorovano učenje, ocenjevanje zanesljivosti, točkovne cenilke, intervalne cenilke.

University *of Ljubljana*
Faculty *of Computer and Information Science*
Darko Pevec
*Reliability estimation of individual predictions in supervised learning*

# ABSTRACT

The thesis discuses reliability estimation of individual predictions in the supervised learning framework. This research field is rather new and currently attracts little attention from experts and users alike. The main indicators of success of machine learning models are consequently still measures of average performance, for example classification accuracy or root mean square error. However, averaged statistics are not able to provide a full view of models' performance and an increasing number of today's users of machine learning have interest for additional information that can help them to better understand the models' results. This additional information is even more important in cases where wrong predictions may lead to serious financial losses or medical complications. It happens that experts become reluctant to use prediction systems if their predictions are not backed up by their reliability assessments. The prevalent methods for model assessment are insufficient among fields where decision support is of crucial importance or where the average performance is not of paramount importance, information on the reliability of single predictions may prove very beneficial.

The greatest concern with use of machine learning algorithms is whether the chosen model represents the data well and if the predicted values conform to the dataset or has the model learned a wrong concept or even over-fitted to noise in the data. As we want to take into account all possible machine learning models, we have to deal with them as with black-boxes, which means we only have access to their input (the training examples) and their output (their predictions).

This work presents a complete overview of reliability estimators for supervised learning. This framework consists of classification and regression, due to their inherent differences. We also distinguish between point-wise and interval estimators, but interestingly, point-wise estimators can be applied both to classification and regression, whereas interval estimators are defined only for regression.

The first contribution of this thesis is a new comparative study of the usefulness of point-wise estimators in the classification setting. The analysis and comparison with a reference function shows that this kind of reliability estimation is rarely useful on real-world datasets. But in cases when we have to deal with a suboptimal model and the point-wise estimators conform with the data, they can prove to improve the results and provide additional information.

Regarding interval estimation, the thesis contributes a novel, unifying view of reliability estimation enabling their comparison, which was not possible before. Our analysis shows the dual nature of the two families of approaches: methods based on bootstrap and maximum likelihood estimation provide valid prediction intervals and methods based on local neighborhoods provide optimal prediction intervals. Based on this finding, we present a combined approach that merges the properties of the two groups. Results of this method are favorable, indicating that the combined prediction intervals are more robust.

Existing statistics that provide information merely on the models' average accuracy are not truly informative, while on the other hand, appropriate graphic visualizations are known to be very useful for developing users' intuition and understanding of the models behavior. After demonstrating an existing visualization tool for comparing prediction intervals we present a new visualization technique that enables model comparison and has potential for knowledge discovery.

The final contribution is a model aggregation procedure based on a combined statistic for robust selection and merging of regression predictions. This new evaluation statistic and aggregation procedure provides confirmatory and consequently more reliable predictions.

*Key words:* supervised learning, reliability estimation, point-wise estimates, interval estimates.

# ZAHVALA

*Največja zahvala gre mentorju, prof. dr. Igorju Kononenku, za vodstvo in usmerjanje pri raziskovalnem delu preteklih nekaj let, iz katerega je nastala tudi pričujoča disertacija. Hvaležen sem za obilico časa, katerega mi je namenil, ter za vse priložnosti spoznavanja novih raziskovalcev širom sveta. Zahvaljujem se doc. dr. Zoranu Bosniću, za ves čas ki sem mu ga ukradel, saj je bilo njegovo delo začetno izhodišče mojih raziskav. Zdaj že doc. dr. Erik Štrumbelj je izrazil veliko zanimanja in mi podaril kar nekaj časa, skupaj sva prebrodila skozi marsikatera vprašanja in tudi našla nekatere nove odgovore. Zahvaljujem se tudi vsem ostalim trenutnim in nekdanjim sodelavcem Laboratorija za kognitivno modeliranje, ki so se pogosto kar mimogrede znašli v diskusijah o mojem delu, za vsa njihova razmišljanja, komentarje in nove ideje. Pri zaključnem delu se posebej zahvaljujem prof. dr. Nadi Lavrač in izr. prof. dr. Marku Robnik-Šikonji, za odlične pripombe in predloge kako izboljšati disertacijo ter za kopico novih idej nadaljnjih raziskav. Zahvala gre tudi doc. dr. Iztoku Lebarju Bajcu, za vso pomoč pri zaključnem oblikovanju. Ne nazadnje se zahvaljujem vsem tistim, ki so me pri tej avanturi podpirali, vedrili in mi stali ob strani, zlasti staršem in samooklicani podporni skupini. Zahvaliti pa se moram tudi Agenciji za raziskovalno dejavnost RS, ki je delo finančno podprla.*

— Darko Pevec, Ljubljana, junij 2013.

# KAZALO

I

# *Uvod*

Za strojno učenje velja, da v širši razvrstitvi spada večinsko na področje umetne inteligence. Kljub temu se dotika še drugih področij, tako računalništva kot matematike in statistike, pa tudi psihologije, filozofije in drugih znanosti [1]. V svetu strojnega učenja so redka dela, ki ne citirajo ravno omenjene temeljne knjige tega področja in je ena izmed prvih knjig, ki nudi obsežen pregled temeljnega znanja.

Skozi leta je področje cvetelo, porajale so se nove paradigme, nastajale in reševale so se nove vrste problemov. Tudi poimenovanja pojmov so se skozi čas spreminjala. Vendar tudi v novejših pregledih področja, na primer [2], ni moč najti dobrega odgovora na precej preprosto, če ne celo nekoliko naivno vprašanje: *so rezultati strojnega učenja res dobri?*. Hitro je moč najti nasvet, da je potrebno minimizirati srednjo kvadratno napako, vendar, ali je to prav? Če primerjamo posamezne napovedi različnih modelov med seboj, gotovo ne, ne malo krat pa se izkaže, da se model z najmanjšo povprečno napako preveč prilega šumu v podatkih.

Postavimo se v vlogo nekoga, ki se prvič srečuje z računalniškim napovedovanjem, s strojnim učenjem. Ko prebrodi skozi začetna vprašanja izbire atributov, števila zbranih podatkov, in sorodna, pride do končnega vprašanja ali bi nek model bil bolj ustrezen kot drugi in kaj je zares najbolje? Obstaja veliko vprašanj in nobenega preprostega celostnega odgovora. Kot bomo videli v razdelku 2.3, obstaja veliko delnih odgovorov in specializiranih rešitev, vendar odgovora na splošno vprašanje ni moč najti. Disertacija se trudi na to vprašanje čim bolje odgovoriti in ponuditi praktična orodja (v dodatku pa tudi napotke), ki lahko mnogim koristijo.

## 1.1 Motivacija

Ocenjevanje zanesljivosti je relativno mlado področje strojnega učenja, trenutno pa mu strokovnjaki in uporabniki ne namenjajo dovolj pozornosti. Na žalost so še zmeraj glavni kazalci uspešnosti modelov strojnega učenja mere povprečnega delovanja, na primer povprečna točnost. Vendar povprečne mere še zdaleč ne nudijo popolne slike in za vse več današnjih uporabnikov metod strojnega učenja so, poleg samih napovedi, zanimive še druge informacije, ki lahko uporabniku pomagajo pri razumevanju modelov. Dodatne informacije so zlasti dobrodošle, če ne kar pomembne, kadar imajo napačne napovedi lahko hude ekonomske ali zdravstvene posledice. Takrat strokovnjaki napovednim sistemom ne zaupajo zlahka, če jim le-ti ne nudijo dovolj informacij o zanesljivosti posameznih napovedi [3]. Ustaljene metode ocenjevanja točnosti napovednih sistemov so nezadostne zato, ker ne nudijo potrebne podpore in dodatnih

informacij o zanesljivosti posameznih napovedi. Na področjih, kjer je podpora pri odločanju pomembna oziroma kjer povprečna uspešnost ni najpomembnejša, bi informacije o zanesljivosti posameznih napovedi bile zelo koristne.

## 1.2 Prispevki k znanosti

Poglavitni prispevki doktorske disertacije k znanosti so:

- KRITIČNA ANALIZA TOČKOVNIH CENILK; v razdelku 4.6 je predstavljena nova primerjalna študija, v kateri raziskujemo ali so točkovne ocene zanesljivosti dobri indikatorji zanesljivosti posameznih napovedi na klasifikacijskih primerih. Analiza z uporabo referenčne ocene, ki temelji zgolj na napovedih modelov (razdelek 4.2), pokaže, da točkovne ocene zanesljivosti le redko izboljšajo napovedi samih modelov;

- naprednejši pristopi, temelječi na intervalnih ocenah:

  CENILKA KI ZDRUŽUJE PRAVILNE IN OPTIMALNE INTERVALNE OCENE; v preteklosti ni bilo opravljene primerjave intervalnih cenilk in naše delo jih je postavilo na skupni imenovalec ter s tem omogočilo njihovo primerjavo. Naša analiza je razkrila dualno naravo intervalnega ocenjevanja, v smislu da ena družina metod optimizira pravilnost in druga optimalnost (mere uspešnosti intervalnih ocen so predstavljene v razdelku 6.2). Kombinirana metoda združuje dve najboljši metodi iz različnih družin in je predstavljena v razdelku 7.1;

  RAZVOJ INFORMATIVNE PRIMERJALNE VIZUALIZACIJE obstoječe statistike, ki nudijo informacije o povprečni uspešnosti modelov, so nezadovoljive. Zaradi tega je v razdelku 7.2.2 predstavljena izboljšana metoda vizualizacije residualov proti napovedanim vrednostim, ki vključuje tudi napovedne intervale in omogoča vizualno primerjavo modelov;

  PREDLOG POSTOPKA, OZIROMA STATISTIKE, S KATERIMI JE MOČ ROBUSTNO IZBIRATI IN ZDRUŽEVATI REGRESIJSKE NAPOVEDI; Na osnovi statistik uspešnosti intervalnih ocen predlagamo v razdelku 7.4 novi način ocenjevanja modelov na osnovi doseženih statistik pravilnosti in optimalnosti, ki omogoča izbiro in združevanje tistih napovedi, ki so najbolj kredibilne, oz. se pravilno prilegajo podatkom.

## 1.3 Pregled naloge

Disertacijo sestavlja osem poglavij in dodatek. Drugo poglavje predstavi širše ozadje in najprej predstavi klasično nadzorovano učenje ter vse modele strojnega učenja, ki smo jih uporabljali pri analizah. Poglavje nudi še podroben pregled širših raziskovalnih področij, povezanih s temo disertacije, skozi pregled obstoječe literature. V tretjem poglavju so predstavljene posplošene interpretacije vseh metod za tvorbo točkovnih ocen zanesljivosti. V četrtem poglavju so opisani poskusi za analizo teh metod, opravljeni na klasifikacijskih problemih. V tem poglavju so predstavljeni tudi tipični primeri obnašanja metod in analiza njihove uporabnosti. Peto poglavje se osredotoča na metode, ki tvorijo intervalne ocene zanesljivosti in šesto poglavje obsega osnovne poskuse z intervalnimi cenilkami na regresijskih problemih. Poleg statistik za vrednotenje intervalnih ocen poglavje nudi vizualizacijo teh metod. V sedmem poglavju so predstavljeni naprednejši pristopi, ki temeljijo na intervalnih ocenah. Poglavje vpelje kombinirano metodo, analizo izbire napovedi z agregacijo modelov in analizo izbire modelov s kombinirano statistiko. Zaključno, osmo poglavje podaja zaključke, razpravo in načrte za nadaljnje delo. Dodatek se osredotoča na praktične vidike ocenjevanja zanesljivosti posameznih napovedi in ponuja praktične napotke za uporabo razvitih metod.

2

# *Ozadje*

V tem poglavju je predstavljeno širše ozadje in temeljni koncepti, potrebni v nadaljnjih poglavjih. Ker je strojno učenje postalo zelo široko področje, se najprej omejimo na, zgodovinsko najstarejše, klasično nadzorovano učenje. Sledi predstavitev različnih uveljavljenih modelov nadzorovanega učenja. Večina ustaljene literature se ukvarja bodisi s klasifikacijo bodisi z regresijo in za razliko od večine ustaljenih razvrstitev je za nas pomembno, kateri modeli so bolj splošni. Zato so posebej ločeni izključno klasifikacijski in izključno regresijski modeli, v tretjo skupino pa so uvrščeni dvotipni modeli, s katerimi je možno reševati oba primera. Zadnji razdelek je namenjen ožjemu pregledu področja oziroma objavljenih del, vezanih na temo disertacije.

## 2.1 Nadzorovano učenje

V začetku poglavja smo že omenili klasifikacijo in regresijo. Oba termina skupaj tvorita klasično nadzorovano učenje. Gre za učenje na poljubnih podatkih, ki so vzorčeni iz neke populacije, načeloma neodvisno. Podatke predstavljajo zajeti primeri, opisani z vektorji vrednosti atributov. Tipična naloga za nadzorovano strojno učenje se pojavi, ko se zbere podatke in si izmed vseh atributov izberemo enega, ga proglasimo za odvisnega, in ga želimo napovedati za nove podatke, kjer je vrednost odvisnega atributa neznana. V primeru, ko je odvisni atribut diskreten (ima končno zalogo vrednosti), imamo opravka s klasifikacijo, in v primeru, ko je ta zvezen (ima neskončno zalogo vrednosti), z regresijo. V splošnem imamo pri učenju lahko opravka tudi z več odvisnimi atributi, zaradi preprostosti pa se omejujemo na samo enega.

Pri uporabi metod strojnega učenja se je potrebno vprašati ali ima izbrani model dobro prestavitev podatkov in ali so napovedane vrednosti konformne, saj se modeli zlahka naučijo napačnih konceptov ali pa se preveč prilagodijo šumu. Ker želimo zaobjeti vse modele strojnega učenja, moramo nanje gledati iz zelo splošnega vidika – kot na črne škatle, za katere poznamo le njihov vhod (učni podatki) in njihov izhod ob poljubnih podatkih (testni podatki, novi primeri). Zaradi te omejitve ne poznamo in nam tudi ni potrebno vedeti ničesar drugega o samem modelu, še najmanj pa kako ta zares deluje. Kar imamo na voljo, je množica opazovanj $(\vec{x_i}, y_i), i = 1..n$, kjer so $\vec{x_i}$ vektorji diskretnih ali zveznih atributov in $y_i$ so zabeležene vrednosti odvisne spremenljivke (atributa), katere se želimo naučiti in kasneje napovedovati za nove primere. Najpreprostejša predpostavka, na katero se zanašamo, je neodvisno vzorčenje, zato da imamo vsaj neko garancijo reprezentativnosti učnih podatkov. V nadaljevanju uporabljamo odvisno spremenljivko tudi za izračune ocen zanesljivosti. Omeniti velja, da

je to pri regresiji realna spremenljivka, pri klasifikaciji pa uporabljamo verjetnostno distribucijo po razredih.

## 2.2 Modeli nadzorovanega učenja

Modele nadzorovanega učenja lahko razvrstimo v tri skupine glede na to, kakšno odvisno spremenljivko znajo obravnavati. Zato jih uvrščamo med klasifikacijske, regresijske in dvotipne modele; s slednjimi je možno reševati oba problema. V nadaljevanju se le bežno dotaknemo modelov, ki jih uporabljamo za testiranja, navajamo pa uporabljeno parametrizacijo.

Vsi poskusi in metode so implementirani v okolju za statistično programiranje R [4]. Okolje je že dolgo uveljavljeno v statističnih krogih. Primerno je tudi za strojno učenje, o čemer pričajo številne implementacije orodij in modelov strojnega učenja.

### 2.2.1 Klasifikacijski modeli

Klasifikacijske modele določa omejitev, da je odvisna spremenljivka lahko zgolj diskretna, kar pomeni, da vrednost odvisnega atributa vsakega primera pripada nekemu razredu iz končno velikega nabora možnih vrednosti.

#### *Naivni Bayes (nb)*

Naloga Bayesovega klasifikatorja je izračunati pogojne verjetnosti za vsak razred pri danih vrednostih atributov, za dani novi primer, ki ga želimo klasificirati. Naivni Bayesov klasifikator pri danem razredu predpostavi medsebojno pogojno neodvisnost atributov. To med drugim omogoča, da učna množica po navadi zadošča za zanesljivo oceno vseh potrebnih verjetnosti za izračun končne pogojne verjetnosti vsakega razreda. Osnovna formula je izpeljana s pomočjo Bayesovega pravila. Naloga učnega algoritma je potem s pomočjo učne množice aproksimirati apriorne in pogojne verjetnosti možnih razredov pri danih vrednostih atributov.

V okolju R najdemo implementacijo naivnega Bayesovega klasifikatorja v funkciji `naiveBayes` paketa `e1071`, za pravilno delovanje pa potrebujemo še paket `class`. Verjetnostno distribucijo dobimo tako, da klasifikacijskemu algoritmu `predict` podamo parameter `type="raw"`. Ker naivni Bayesov klasifikator računa pogojne verjetnosti diskretnih spremenljivk, je potrebna diskretizacija zveznih atributov, ki je v uporabljeni implementaciji že vgrajena.

*Odločitvena drevesa (od)*

Odločitveno drevo [5] je sestavljeno iz notranjih vozlišč, ki ustrezajo atributom, vej, ki ustrezajo podmnožicam vrednosti atributov, in listov, ki določajo razred. Ena pot v drevesu od korena do lista ustreza enemu odločitvenemu pravilu. Pri tem so pogoji, tj. pari atribut–podmnožica vrednosti, ki jih obiščemo na poti, konjuktivno povezani. Algoritmi odločitvenih dreves za gradnjo uporabljajo ocene informativnosti za izbiro posameznih atributov in ustreznih podmnožic njihovih vrednosti.

V okolju za statistično programiranje R najdemo implementacijo odločitvenih dreves v paketu `rpart`. Verjetnostno distribucijo dobimo tako, da učnemu algoritmu podamo parameter `method="class"`.

### 2.2.2 Regresijski modeli

Regresijske modele določa zvezna odvisna spremenljivka, kar pomeni, da je (teoretično gledano) zaloga vrednosti odvisne spremenljivke neskončna.

*Linearna regresija (lr)*

Linearna regresija [6] je zgodovinsko prva statistična metoda, ki modelira odvisnost med odvisno spremenljivko $y_i$ in neodvisnimi spremenljivkami $\vec{x}_i$ kot linearno enačbo

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_a \cdot x_{ia} + \epsilon_i,$$

kjer so $\beta_.$ koeficienti, $\epsilon_i$ pa nerazložena napaka. Koeficiente se izračuna iz predoločenega sistema linearnih enačb, tako da se minimizira skupna napaka, za katero se predpostavlja normalna porazdelitev. V praksi se ta sistem enačb rešuje v matrični obliki z metodo singularne dekompozicije [7], ki odstrani vpliv manj pomembnih spremenljivk in tako zagotovi preprostejši in posledično razumljivejši model. V okolju R najdemo implementacijo linearne regresije v osnovni distribuciji, pod oznako `lr`.

*Regresijska drevesa (rd)*

Model regresijskih dreves [5] je skoraj identičen odločitvenim drevesom. Bistvena razlika je v tem, da regresijska drevesa v listih uporabljajo neko obliko regresije, po navadi izvajajo kar linearno regresijo ali pa povprečje primerov v listih. V okolju R najdemo implementacijo regresijskih dreves v istem paketu kot odločitvena drevesa, v paketu `rpart`. V listih je minimalno en primer, minimalno število primerov za poskus delitve je 2 in parameter kompleksnosti pri obrezovanju je 0.01.

### 2.2.3 Dvotipni modeli

Med dvotipne modele uvrščamo tiste modele, katerim je načeloma vseeno, ali opravljajo klasifikacijo, ali regresijo. V njihovi implementaciji ni bistveno, ali je odvisna spremenljivka diskretna ali zvezna in so s tega stališča dvotipni.

#### *Najbližji sosedi (11ns)*

Najpreprostejša varianta algoritma najbližjih sosedov kot znanje uporablja kar množico vseh učnih primerov, učni algoritem si zgolj zapomni vse primere. Pri določanju odvisne spremenljivke novega primera se iz učne množice poišče določeno število najbolj podobnih, tj. najbližjih primerov – sosedov v atributnem prostoru. Odvisno spremenljivko novega primera določa povprečje bližnjih sosedov. Zaradi ustrezne metrike v prostoru atributov je pri tem potrebno normalizirati vrednosti zveznih atributov in definirati razdaljo (metriko) med vrednostmi vsakega diskretnega atributa. Pri testiranjih smo uporabljali hitro implementacijo s prekrivnimi drevesi, `knnx` paketa `FNN`, v kateri smo uporabljali konstantno velikost okolice 11 sosedov, kar bistveno izniči vpliv šuma.

#### *Metoda podpornih vektorjev (mpv)*

Naloga tega učnega algoritma je izračunati vrednosti koeficientov vnaprej podane funkcije, ki predpostavlja hiperploskev v atributnem prostoru [8]. Hiperravnino lahko definiramo samo v zveznem prostoru, torej morajo biti vsi atributi zvezni. Če imamo opravka z več razredi, potrebujemo za vsak par eno hiperravnino. Uporabljene funkcije so lahko linearne, kvadratične, polinomske, praktično poljubne, le da izpolnjujejo pogoje notranjega produkta (jedrni trik). Koeficiente hiperravnine se iterativno določa tako, da se napaka postopoma zmanjšuje.

Implementacija metode `svm` se v okolju R nahaja v paketu `e1071`. Ker se ne ukvarjamo z optimizacijo, smo vse nastavitve prepustili avtorjem paketa, oziroma uporabili privzete vrednosti.

#### *Umetna nevronska mreža (unm)*

V nadzorovanem učenju se najpogosteje uporablja usmerjene večnivojske umetne nevronske mreže [9]. Te kaskadno povezujejo več nivojev nevronov: vhodni nevroni (ki ustrezajo atributom), eden ali več nivojev skritih nevronov in izhodni nevroni (ki ustrezajo zalogi vrednosti odvisnega atributa). Naloga učnega algoritma je nastaviti uteži na povezavah med nevroni tako, da je napaka čim manjša.

Za novi primer dobi nevronska mreža na vhodne nevrone vrednosti atributov. Zatem vsak nevron na naslednjem nivoju izračuna svoj izhod kot uteženo vsoto svojih vhodnih signalov, ki jo nelinearno normalizira.

V okolju R so umetne nevronske mreže z enim skritim nivojem implementirane v paketu `nnet` [10]. Pri testiranjih uporabljamo povprečne, hevristične nastavitve; nastavili smo število nevronov skrite plasti s parametrom `size=5`, uporabljamo klasično vzvratno razširjanje, linearne funkcije, preskočne povezave, algoritem pa smo omejili na 1000 iteracij. Da za primer klasifikacije dobimo verjetnostno distribucijo, algoritmu dodatno podamo parameter `type="raw"`.

### *Mreža radialnih baznih funkcij (mrbf)*

Mrežo radialnih baznih funkcij [11] lahko umestimo v širšo družino umetnih nevronskih mrež, od že predstavljenih pa se razlikuje v dveh bistvenih točkah. Mreža uporablja radialne bazne funkcije, kar v našem primeru pomeni, da vsak nevron predstavlja Gaussovo distribucijo. Druga razlika tiči v strukturi. V naši implementaciji pri testiranjih uporabljamo toliko baznih nevronov, kolikor je učnih primerov, 2 najbližja soseda pa določata varianco tj. raztros posamezne bazne funkcije.

### *Bagging (bag)*

Izraz bagging pride iz izraza "'bootstrap agreggating"' [12]. Stremljenje (angl. *bootstraping*) je bolj splošen princip razmnoževanja učnih primerov, kadar jih nimamo dovolj za učenje, pri baggingu pa generiramo serijo različnih učnih množic. Obstajajo primeri, v katerih metoda boosting [13] dosega višjo točnost kot metoda bagging, vendar pri njej obstaja večja verjetnost prekomernega prileganja šumu v podatkih.

Če ima učna množica $n$ primerov, potem vsakič $n$ krat naključno izberemo primer iz učne množice z vračanjem. To pomeni, da se isti učni primer v tako vzorčeni učni množici lahko večkrat ponovi, nekaterih primerov iz originalne učne množice pa generirana množica sploh ne vsebuje. Nad vsako generirano učno množico zatem poženemo učni algoritem in tako dobimo veliko število različnih hipotez. Množico generiranih hipotez potem uporabimo za napovedovanje novega primera s povprečenjem napovedi vseh hipotez. Pri klasifikacijskih problemih uporabljamo odločitvena drevesa in regresijska drevesa pri regresijskih problemih.

Implementacijo metode `bagging` najdemo znotraj okolja R v paketu `ipred`. Za rezultat v obliki verjetnostne distribucije pri uporabi odločitvenih dreves je potrebno

algoritmu podati parameter `type='prob'`.

### *Naključni gozdovi (ng)*

Metoda naključnih gozdov je namenjena izboljševanju napovedne točnosti drevesnih algoritmov. Originalno je bila razvita za odločitvena drevesa [14]. Ideja je generirati množico dreves, tako da se pri izbiri najboljšega atributa v vsakem vozlišču naključno izbere relativno majhno število atributov, ki vstopajo v ožji izbor za najboljši atribut. Vsako tako zgrajeno drevo se zatem uporablja za napovedovanje novih primerov po metodi glasovanja.

V okolju R je metoda naključnih gozdov implementirana v paketu `randomForest`. Za rezultate v obliki verjetnostnih distribucij je potrebno algoritmu podati parameter `type='prob'`, enako kot pri metodi bagging.

## 2.3 *Pregled področja*

Primeren kriterij za razlikovanje med mnogimi različnimi pristopi k ocenjevanju zanesljivosti posameznih napovedi je, ali so namenjeni specifičnim modelom ali pa so neodvisni od modelov. Mnogi raziskovalci razvijajo pristope za umetne nevronske mreže, za nas pa so predvsem zanimivi pristopi, ki so neodvisni od modelov. Za slednje po navadi pravimo, da delujejo po principu *črnih škatel*, s čemer izpostavljamo, da ne vemo ničesar o modelih — vemo, kaj dobijo na vhod in kaj vrnejo na izhod, kaj se dogaja znotraj modela, pa nam je nedosegljivo in nas posledično ne zanima. Ti pristopi temeljijo na izkoriščanju osnovnih gradnikov nadzorovanega učenja, to so učni primeri in njihovi atributi. Raziskave so tesno povezane s področjem meta učenja, ki prav tako poskuša ugotoviti močne in šibke točke posameznih modelov.

Skupina pristopov, ki izboljšujejo točnost napovednih modelov s pomočjo vzorčenja, so splošni in neodvisni od modelov. Eden izmed najbolj znanih takšnih pristopov je metoda bagging [12], katero smo obravnavali v razdelku 2.2.3, ki temelji na vzorčenju učne množice z vračanjem. Med podobne pristope združevanja modelov sodi metoda *boosting* [13], ki temelji na sekvenčnem učenju šibkih klasifikatorjev. Primeri, pri katerih je napaka večja, dobijo v naslednji iteraciji učenja večjo utež. Končni model predstavlja uteženo povprečje vseh šibkih klasifikatorjev. Izkazalo se je, da takšno združevanje učinkovito izboljša pristranskost in varianco napovedi, zmanjša pa tudi celokupno napako. Raziskave so pokazale, da so šibki klasifikatorji lahko zelo preprosti, z uporabe omenjenega postopka pa so skupaj zmožni zajeti kompleksne odločitvene

meje. V kasnejših delih je bil postopek preveden na regresijske probleme [15]. Zelo sorodna pristopa, ki ravno tako temeljita na uporabi več modelov in združevanju njihovih napovedi, sta *stacking* [16] in *bumping* [17].

Postopek *dvojne perturbacije in kombinacije* (angl. *dual perturb and combine*) [18] je primer pristopa, ki temelji na perturbaciji testnih primerov. Na prvem koraku se na osnovi učnih podatkov zgradi model, ki se tekom postopka ne spreminja. Pri napovedovanju testnih primerov se atributni vektor večkrat perturbira z aditivnim šumom. Dobljene napovedi se nato združi s povprečenjem, kar v končni fazi nudi bolj stabilno napoved od originalnega modela. Poskusi so pokazali, da metoda lahko v določenih primerih tekmuje s pristopom bagging. Ima tudi to prednost, da ne generira večje množice modelov, kar je časovno učinkovitejše, hkrati pa ohranja enostavno razlago uporabljenega modela. Pristop stremljenja je bil uporabljen tudi pri nenadzorovanem učenju, za ocenjevanje zanesljivosti razvrščanja v skupine [19]. Avtorji so generirali večje število razvrstitev in analizirali frekvence, kolikokrat je bil kateri primer razvrščen v katero skupino, povprečne vrednosti pa nudijo bolj stabilno razvrščanje.

V primerih, ko so na voljo večje količine neoznačenih podatkov (znani so le atributni vektorji), lahko njihova uporaba v navezavi z označenimi podatki bistveno izboljša napovedno moč napovednega modela [20]. Neoznačeni primeri ne nudijo direktnega znanja o skriti zvezi med atributi in odvisno spremenljivko temveč tekom učenja ponudijo dodatne informacije o distribuciji prave vrednosti v atributnem prostoru, kar posledično doprinese k večji točnosti učenja. Za tem stoji znani postopek maksimizacije pričakovanja (angl. *expectation maximization*) [21]. Neoznačeni podatki so pogosto na voljo, njihovo označevanje pa je lahko drago in/ali časovno zahtevno; pogosto najdemo takšne primere pri medicinskih podatkih. Najdemo jih tudi pri razpoznavanju slik in adaptacija klasifikatorjev na tem področju tudi izboljšuje rezultate [22].

Uporaba neoznačenih primerov je zanimiva tudi v kontekstu ko-učenja, kjer je vsak učni primer sestavljen iz dveh neodvisnih in vzajemno redundantnih delov [23]. Na predpostavki, da je vsak izmed neodvisnih delov zadosten za uspešno učenje, so avtorji pokazali, da je mogoče uporabljati model, naučen na enem delu podatkov, za označevanje neoznačenih primerov drugega dela podatkov. Podoben pristop najdemo v [24], kjer gre za poskus učenja iz popolnoma neoznačenih podatkov – oblika samo-učenja. Čeprav doseženi rezultati niso blesteči, je študija zanimiva, saj predstavlja inovativen pristop k učenju neoznačenih podatkov.

V raznih raziskavah na temo zanesljivosti se pogosto srečamo s terminom *transdukcije*

in *transduktivnega sklepanja*. Gre za princip sklepanja, s posameznega na posamezno [25]. V našem kontekstu to na primer pomeni, da o neoznačenih primerih sklepamo iz dosegljivih označenih primerov, brez učenja splošnega modela. Ker je princip zelo splošen, obstaja tudi v kontekstu zanesljivosti veliko pristopov, na primer [26, 27] in [28, 29].

Zanimivo področje raziskav se ukvarja z *analizo senzitivnosti*, pri čemer se opazuje vpliv parametrov modelov in drugih lastnosti na njihovo strukturo in napovedi [30]. Analizo se po navadi izvaja kot niz poskusov, pri katerih se sistematično spreminja določene vhodne parametre in beleži njihov vpliv na spremembe izhoda. Ker je pristop neodvisen od modela, so ga uporabili na mnogih področjih. Primere iz strojnega učenja najdemo na umetnih nevronskih mrežah [31] in Bayesovskih mrežah [32]. Analize stabilnosti različnih učnih modelov na osnovi empirične napake in metode testiranja *izpusti-enega* najdemo v delih [33, 34].

Teoretični zametki intervalnega ocenjevanja segajo v začetke moderne statistike. Prvo delo s tega področja je [35], v katerem so predstavljeni intervali zaupanja. V skupnosti strojnega učenja je bil problem ocenjevanja variance distribucije odvisne spremenljivke, ko za šum ni predpostavljena enakomerna porazdelitev v atributnem prostoru, obravnavan v [36]. Njihova metoda razširja umetno nevronsko mrežo z dodatnim izhodnim nevronom, ki je zadolžen za izračun lokalne variance odvisne spremenljivke in nudi informacijo o negotovosti posameznih napovedi.

Svoj čas je obstajalo več statističnih pristopov k oblikovanju intervalov zaupanja za homoskedastične podatke. V [37] najdemo primerjavo treh takšnih pristopov (metoda delte, stremljenje in metoda sendviča) na modelu umetnih nevronskih mrež. Testiranja so pokazala, da stremljenje nudi najbolj točne ocene standardizirane napake napovedanih vrednosti.

Tudi pri intervalnem ocenjevanju zanesljivosti obstaja mnogo pristopov, ki so prikrojeni posameznim modelom in večina literature prihaja s področja umetnih nevronskih mrež. Medtem ko se intervali zaupanja ukvarjajo z deviacijami napovedi modelov od pričakovane pogojne vrednosti, se napovedni intervali osredotočajo na razlike med napovedmi in pravimi vrednostmi. V prvi objavi, v kateri se pojavijo napovedni intervali [38], so ti izračunani v dveh korakih. Najprej se izračuna intervale zaupanja za množico umetnih nevronskih mrež, katera nastane s stremljenjem originalne učne množice. Varianco distribucije odvisne spremenljivke pa se izračuna s pomočjo primerov, ki so bili pri vzorčenju izvzeti. Residuali teh primerov so naprej predani posebni

umetni nevronski mreži, ki z uporabo eksponentne aktivacijske funkcije zagotavlja pozitivne ocene variance.

V delu [39] najdemo združitev zadnjih dveh omenjenih del, najdemo pa tudi primerjavo z analitičnimi pristopi. Njihovi rezultati prikažejo jasno premoč kombinacije stremljenja in maksimalnega verjetja pri postavljanju napovednih intervalov nad analitičnimi pristopi.

Popolnoma drugačen pristop računanja napovednih intervalov najdemo v [40], kjer avtorji razvrstijo atributni prostor v mehke skupine. Napovedne intervale določa utežena vsota po pripadnosti posameznim skupinam, napovedni interval posamezne skupine pa določa empirična distribucija residualov znotraj skupine. Pristop je zanimiv zlasti zato, ker se izogne izračunom intervalov zaupanja.

Lokalne okolice so uporabljene v različne namene. V okvirjih klasifikacije obstaja način ocenjevanja podatkovnih množic, ki temelji na stopnji prekrivanja posameznih razredov, ki jo ocenjujejo z *R-vrednostjo* [41]. V predelih atributnega prostora, kjer je R-vrednost visoka, je stopnja prekrivanja višja in tudi testi so pokazali, da je tam točnost napovedi manjša. V preteklih letih je vse manj objavljene literature na temo napovednih intervalov, vendar je vse več raziskav opravljenih na področju kvantilne regresije, katere začetke najdemo v [42]. Obstaja zanimiva reinterpretacija modela naključnega gozda, ki pokaže ekvivalenco med njegovimi napovedmi in uteženim povprečjem opazovanj odvisne spremenljivke [43]. Avtorja sta pokazala, da so naključni gozdovi oblika adaptivnih najbližjih sosedov. Istega leta je bil koncept naključnih gozdov posplošen na Kvantilni Regresijski Gozd [44], ki pa že sam po sebi nudi napovedne intervale.

## 2.4 Vpetost med sorodne raziskave

Na Fakulteti za računalništvo in informatiko, zlasti v Laboratoriju za kognitivno modeliranje, se že vrsto let ukvarjamo med drugim tudi z analizami zanesljivosti posameznih napovedi, ki so posledica bolj zgodnjih raziskav ocenjevanja iz izbora posameznih atributov. Sprva so bile raziskave osredotočene na ocenjevanje zanesljivosti klasifikacij in cenovno občutljivo kombiniranje metod strojnega učenja [28], od koder povzemamo metodo transdukcije, za katero je bilo na praktičnem primeru pokazano da lahko izboljša senzitivnost in hkrati ne spremeni verjetnosti napake prvega reda. Kmalu je sledila analiza metode ocenjevanja zanesljivosti posameznih regresijskih napovedi, ki temelji na principu transdukcije [29]. Sledilo je Ocenjevanje zanesljivosti posameznih napovedi z analizo občutljivosti regresijskih modelov [45], ki med drugim novo me-

todo primerja s klasičnimi in prilagojenimi ocenami zanesljivosti z drugih področij. Posplošitev teh metod tvori poglavje 3, poglavje 4 pa predstavlja komplement poskusov v [45].

3

# *Točkovne cenilke zanesljivosti*

Pristopi, neodvisni od uporabljanih modelov, ne morejo izkoriščati parametrov, specifičnih za te modele. Namesto tega so ti pristopi zasnovani na spreminjanju dosegljivih parametrov klasičnega nadzorovanega učenja — učne množice in atributov. Tako zasnovane cenilke zanesljivosti nudijo ocene v obliki metrik nad možnimi opazovanimi lastnostmi.

Ker so metode zasnovane na hevurističnih interpretacijah dosegljivih podatkov, lahko te metrike vračajo rezultate iz poljubnega intervala in kot takšne nimajo verjetnostne razlage. Metode po naravi merijo odklon od želene točne napovedi in so v resnici cenilke nezanesljivosti. V primeru regresije te cenilke vrnejo poljubna pozitivna realna števila (z intervala [0,∞]), tako da 0 predstavlja najbolj zanesljivo napoved, ostale vrednosti pa predstavljajo vse večje stopnje nezanesljivosti. Če želimo bolj naravno interpretacijo, da je večje boljše, je za regresijo možna transformacija obratna vrednost (zanesljivost=1/nezanesljivost), tako da je 0 najmanj zanesljivo in pozitivne vrednosti so bolj zanesljive. V primeru klasifikacije pa lahko zaradi končno velike napake (z intervala [0,1]) ocene zanesljivosti preprosto preoblikujemo v zanesljivost (zanesljivost=1-nezanesljivost), tako da 0 predstavlja najbolj nezanesljivo in 1 najbolj zanesljivo napoved.

Nadaljevanje poglavja nudi pregled pristopov za tvorbo točkovnih ocen zanesljivosti. V delu [45] najdemo prvi pregled ocen zanesljivosti v regresiji in primerjalno analizo različnih mer zanesljivosti. Prevedbo teh metod na problem klasifikacije je moč najti v [46]. Tukaj nudimo posplošeni pogled na te metode z vidika nadzorovanega učenja, podobno kot v [47].

Pri ocenah časovnih zahtevnosti uporabljamo klasično $\mathcal{O}$ notacijo, kjer $n$ predstavlja število razpoložljivih učnih primerov, $m$ število atributov, za model $M$ pa z $\mathcal{O}(M)$ označujemo časovno zahtevnost napovednega algoritma in z $\mathcal{O}(M_n)$ časovno zahtevnost učnega algoritma.

## 3.1 Lokalno modeliranje napake napovedi

Ta pristop točkovnega ocenjevanja zanesljivosti posameznih napovedi temelji na oznakah najbližjih sosedov posameznega primera. Ko za posamezni novi primer dobimo od poljubnega modela napoved $K$, poiščemo množico $k$ najbližjih sosedov v učni množici $\left[(\vec{x_1}, C_1), (\vec{x_2}, C_2), \ldots, (\vec{x_k}, C_k)\right]$, kjer so prave oznake odvisne spremenljivke na teh sosedih označene s $C_i$. Pri testiranjih smo uporabljali okolico velikosti 5. Cenilka, ki lokalno modelira napako napovedi, je definirana kot povprečna razdalja med oznakami

$k$ sosedov in napovedano vrednostjo $K$:

$$CNK = \frac{\sum_{i=1}^{k} \|C_i, K\|}{k}. \tag{3.1}$$

Oznaka $CNK$ pride iz angleškega zapisa $C_{Neighbors} - K$. V primeru klasifikacije lahko načeloma izbiramo iz množice različnih mer razdalj, pri regresiji pa je najbolj naravna kar evklidska razdalja. Pri regresiji lahko zaradi obravnave zveznih vrednosti posebej ločimo $CNK_a$, ki označuje absolutno vrednost cenilke, in $CNK_s$, ki označuje predznačeno oceno.

Časovna zahtevnost je najbolj odvisna od časa, potrebnega za iskanje najbližjih sosedov. Uporaba prekrivnih dreves [48] omogoča iskanje najbližjih sosedov v $\mathcal{O}(m \cdot \log n)$. Za eno oceno $CNK$ potrebujemo eno napoved modela, $k$ iskanj sosedov in malenkost aritmetike (reda števila razredov). Potrebujemo tudi predpripravljeno prekrivno drevo, ki se zgradi v $\mathcal{O}(n \cdot \log n)$ času. Tako je časovna zahtevnost

$$\begin{aligned} \mathcal{O}(CNK(\vec{x})) &= \mathcal{O}(k \cdot (m \cdot \log n + 1) + M) \\ &= \mathcal{O}(m \cdot \log n + M). \end{aligned} \tag{3.2}$$

## 3.2 Lokalno prečno preverjanje

Če prečno preverjanje izvedemo nad manjšim, okoliškim delom vhodnega podprostora, ta nudi lokalno oceno napovedne napake, torej lahko služi kot ocena zanesljivosti. Za novi, neoznačeni primer poteka izračun ocene $LCV$ (iz angleško *local cross-validation*) na naslednji način. Najprej poiščemo lokalno okolico novega primera, t.j. množico $k$ najbližjih sosedov. Velikost te množice je parameter metode, ki se ga po navadi izraža z odstotki velikosti učne množice. V [45] je ta parameter 5%, v [49] in tukaj pa pri preizkusih uporabljamo 10% učnih primerov. To je bistveno več kot 10 učnih primerov, kolikor se v splošnem uporablja pri klasifikaciji, vendar je gradnja kompleksnih modelih na tako malem številu primerov vprašljiva. Uporaba večje množice, izražene v odstotkih, zagotavlja večjo stopnjo robustnosti.

Generirati moramo $k$ novih modelov, pri katerih je vsakič izpuščen eden izmed $k$ sosedov. S temi novimi modeli začnemo graditi novi pogled na okolico novega primera tako, da najprej zabeležimo modelove napovedi izpuščenih primerov; z $i$-tim novim modelom napovemo $K_i$. Ker so ti primeri v učni množici, poznamo njihove prave oznake $C_i$. Zdaj lahko zapišemo oceno zanesljivosti z uporabo lokalnega prečnega

preverjanja kot povprečno razdaljo (razliko) med bližnjimi napovedanimi vrednostmi in dejanskimi:

$$LCV = \frac{1}{k}\sum_{i=1}^{k} \|C_i, K_i\| . \tag{3.3}$$

Kot vidimo, pri tej cenilki sama napoved novega primera $K$ ne igra vloge pri ocenjevanju zanesljivosti, temveč le napovedi njegove lokalne okolice. Parameter $k$ je linearno odvisen od števila primerov, zato je časovna zahtevnost izračuna ocene

$$\begin{aligned} \mathcal{O}(LCV(\vec{x})) &= \mathcal{O}(k \cdot (m \cdot \log n + M_k + M + 1)) \\ &= \mathcal{O}(n \cdot (m \cdot \log n + M_n + M)). \end{aligned} \tag{3.4}$$

## 3.3 Varianca modela bagging

Varianca napovedi je bila prvič uporabljena za indirektno ocenjevanje zanesljivosti pri agregiranih nevronskih mrežah, vendar je metodo moč zlahka posplošiti na poljubne modele. Model bagging smo spoznali v razdelku 2.2.3, za točkovne ocene zanesljivosti pa lahko uporabimo varianco napovedi, dobljenih s tem modelom. Naj bo $K$ še naprej napoved novega primera. Ob uporabi množice $b$ modelov zgrajenih z vzorčenjem s ponavljanjem, naj bodo $B_i$, $i = 1 \ldots b$ napovedi teh modelov za novi primer. Ker nas zanima varianca modela bagging, $BAGV$ (iz angleško *BAGging Variance*), je cenilka:

$$BAGV = \frac{1}{m}\sum_{i=1}^{b} (\|B_i, K\|)^2. \tag{3.5}$$

Časovna zahtevnost tega izračuna je

$$\begin{aligned} \mathcal{O}(BAGV(\vec{x})) &= \mathcal{O}(M + b \cdot (M_n + M + 1) + 1) \\ &= \mathcal{O}(b \cdot (M_n + M + 1)). \end{aligned} \tag{3.6}$$

## 3.4 Gostota učnih primerov

Ta pristop sloni na predpostavki, da je za napovedi primerov, ki ležijo v gostejših predelih atributnega prostora, na voljo več informacij in so zato bolj zanesljive. Nasprotno velja v redkejših predelih atributnega prostora, kjer je manj učnih primerov, in tam naj bi napovedi bile manj zanesljive. Ta predpostavka je tipično aplicirana pri odločitvenih in regresijskih drevesih, kjer zaupanje v napovedi narašča sorazmerno s številom

učnih primerov, ki ležijo v istem listu drevesa kot primer, za katerega se napoveduje. Kljub temu je glavna pomanjkljivost tega pristopa v tem, da ne upošteva oznak odvisne spremenljivke. Zaradi tega metoda že apriorno ne more dobro delovati na šumnih podatkih in na primerih, ki niso jasno ločljivi.

Cenilka *DENS* (iz angleško *DENSity*) je ocena gostote učne množice v okolici novega, neoznačenega primera. Ocenjevanje gostote se izvaja s pomočjo Parzenovih oken in z uporabo Gaussove jedrne funkcije. Problem večrazsežnih Gaussovih jedrnih funkcij je moč prevesti na dvorazsežne jedrne funkcije ob uporabi funkcije razdalje med pari atributnih vektorjev. Če označimo učno množico velikosti $n$ kot $\left\{(\vec{x_1}, y_1), (\vec{x_2}, y_2), \ldots, (\vec{x_n}, y_n)\right\}$ in nas zanima ocena gostote za primer $\vec{x}$, to izračunamo z

$$p(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\vec{x}, \vec{x_i}), \tag{3.7}$$

kjer je $\kappa$ Gaussova jedrna funkcija, $p(\vec{x})$ pa vsebuje povprečno obratno oddaljenost do vseh primerov učne množice. Ker je pri gostoti večje boljše, za nezanesljivost pa velja obratno, je za končno oceno gostota obrnjena z maksimumom doseženim na učni množici:

$$DENS = \max_{i=1..n} \left(p(\vec{x_i})\right) - p(\vec{x}). \tag{3.8}$$

Za to oceno moramo izračunati razdalje do vseh učnih primerov in njihove jedrne produkte, tako je časovna zahtevnost za eno oceno

$$\mathcal{O}(DENS(\vec{x})) = \mathcal{O}(n \cdot (m+1)). \tag{3.9}$$

## 3.5 Obratna transdukcija, analiza občutljivosti

Transdukcija pomeni sklepanje s posameznega na posamezno. Ta način sklepanja vključuje sklepanje iz učnih podatkov na odvisno spremenljivko novega primera. Transdukcijo pa lahko izkoristimo v obratni smeri, tako da opazujemo obnašanje modelov (spremembe njihovih napovedi) ob vstavljanju spremenjenih učnih instanc novega, še neoznačenega primera. Za klasifikacijo je bila metoda predstavljena v [28].

Čeprav je transdukcija zelo splošen termin, je za razliko od prej opisanih pristopov ni možno opisati z enim samim postopkom, ki bi zaobjel tako klasifikacijo kot regresijo. Z $\vec{x}$ še zmeraj označujemo vrednosti atributov novega primera. Ne glede na to, ali je problem klasifikacijski ali regresijski, najprej s $K$ zabeležimo napoved modela za $\vec{x}$ na

originalni učni množici. Temu sledi vstavljanje spremenjenih inačic novega primera v učno množico in opazovanje sprememb napovedi na novo naučenih modelov.

V primeru klasifikacije lahko obratno transdukcijo izvedemo na tri različne načine:

- v učno množico vstavimo $\vec{x}$ z enako oznako, kot jo je napovedal model. Ob tem bi moral model le okrepiti svojo napoved, oziroma prepričanje v podano napoved,
- v učno množico vstavimo $\vec{x}$ z oznako, ki je druga najbolj verjetna glede na napoved modela. Ob tem se pojavi vprašanje, ali se bo model zlahka pustil zavesti, oziroma koliko se bo začetna napoved spremenila,
- v učno množico vstavimo $\vec{x}$ z najmanj verjetno oznako, tj. oznako, kateri je model sprva pripisal najmanjšo verjetnost. Tudi tokrat je najzanimivejše vprašanje, koliko se bo spremenila začetna napoved.

Na učni množici, spremenjeni po enem izmed opisanih načinov, ponovno naučimo model. Razdalja med prvotno napovedjo in napovedjo po spremembi učne množice tvori oceno zanesljivosti. Glede na izbran način spreminjanja novega primera, ločimo cenilke $TRANS_{enak}$, $TRANS_{drugi}$ in $TRANS_{zadnji}$.

V primeru regresije govorimo o analizi senzitivnosti. Postopek je podoben, le da se v učno množico iterativno vstavlja množica nekoliko spremenjenih $\vec{x}$. Pristop določa nabor $\epsilon$ vrednosti, ki določajo te majhne spremembe. Postopek je možno zaključiti na dva načina; lahko se osredotočimo na varianco ali pa na pristranskost. Ker v nadaljevanju opisujemo poskuse s točkovnimi ocenami le na klasifikacijskih problemih, so podrobnosti analize senzitivnosti izpuščene – več podrobnosti je moč najti v [3].

Za izračun potrebujemo začetno napoved, spremembo primera, gradnjo novega modela, napoved spremenjenega primera in izračun razlike. Časovna zahtevnost ene ocene je zato

$$\begin{aligned} \mathcal{O}(TRANS(\vec{x})) &= \mathcal{O}(M + M_n + M + 1) \\ &= \mathcal{O}(M + M_n). \end{aligned} \tag{3.10}$$

4

# *Poskusi na klasifikacijskih problemih*

V poglavju so najprej predstavljene testne klasifikacijske množice, katere uporabljamo v nadaljnjih empiričnih poskusih. Točkovne cenilke so bile na regresijskih problemih že preizkušene, kjer se je pristop analize senzitivnosti dobro izkazal [45].

Sledi predstavitev preproste referenčne točkovne cenilke in popravljena metodologija testiranja. V tem poglavju uporabljamo modele iz poglavij 2.2 in 2.2.3, saj izključno regresijski modeli niso primerni za klasifikacijske probleme. Poudariti velja, da so točkovne cenilke zanesljivosti posameznih napovedi dovolj splošne za rabo z poljubnimi modeli nadzorovanega učenja.

Temu sledi analiza vpliva uporabe različnih mer razdalj. Nato sledijo grafični prikazi značilnih situacij. Poglavje zaključujejo prikazi korelacijskih koeficientov in kritična analiza uporabnosti točkovnih ocen zanesljivosti, s poudarkom na empirični analizi sposobnosti izbire najboljših posameznih napovedi.

## 4.1 Testne množice

Za testiranje točkovnih cenilk zanesljivosti v nadaljevanju uporabljamo dvajset realnih oziroma praktičnih testnih množic. Tabela 4.1 prikazuje osnovne lastnosti teh ustaljenih klasifikacijskih podatkovnih množic, prosto dostopnih na repozitoriju strojnega učenja UCI [50]. Iz tabele je moč razbrati, da imamo 5 množic z izključno diskretnimi atributi, 9 množic ima izključno zvezne atribute, medtem ko ima 5 množic mešane atribute. Večina množic ima 2 ali 3 razrede, množica z največ razredov jih ima 10.

## 4.2 Referenčna ocena

Od ocen zanesljivosti pričakujemo, da nam nudijo vpogled v napovedne napake, zato najprej pričakujemo določeno stopnjo pozitivne korelacije med ocenami zanesljivosti in napako. Za bolj formalno definicijo zanesljivosti posamezne napovedi se moramo najprej posvetiti sami napaki posameznih napovedi. Začnimo s posameznim primerom $\vec{x}$, za katerega vemo, da pripada razredu $y$. Naj bo prava pogojna verjetnost $i$-tega razreda $p_i(\vec{x}) = P(Y = i|X = \vec{x})$ in naj $f_i(\vec{x})$ predstavlja napovedano verjetnost tega razreda. V primeru hipotetičnega optimalnega modela bi veljalo $\forall i : f_i(\vec{x}) = p_i(\vec{x})$ oz. $f_y(\vec{x}) = 1$ in 0 za ostale razrede. Tu uporabljamo bolj preprosto, vendar manj pogosto, Laplaceovo napako [51]:

$$e(x) = |y - f(\vec{x})|\,. \tag{4.1}$$

*Tabela 4.1*

Glavne značilnosti testnih množic

| množica | št. primerov | št. diskretnih atributov | št. zveznih atributov | št. razredov |
|---|---|---|---|---|
| housevotes | 435 | 16 | 0 | 2 |
| wine | 178 | 0 | 13 | 3 |
| parkinsons | 195 | 0 | 22 | 2 |
| zoo | 101 | 16 | 0 | 7 |
| tic-tac-toe | 958 | 8 | 0 | 2 |
| postoperative | 90 | 7 | 1 | 3 |
| monks-3 | 432 | 5 | 0 | 2 |
| irisset | 150 | 0 | 4 | 3 |
| glass | 214 | 0 | 9 | 6 |
| hungarian | 294 | 7 | 6 | 2 |
| ecoli | 336 | 0 | 7 | 8 |
| heart | 303 | 7 | 6 | 2 |
| haberman | 306 | 0 | 3 | 2 |
| flag | 194 | 18 | 10 | 10 |
| wdbc | 569 | 0 | 30 | 2 |
| breast-cancer | 369 | 0 | 9 | 2 |
| sonar | 111 | 0 | 60 | 2 |
| hepatitis | 155 | 13 | 6 | 2 |
| lungcancer | 32 | 56 | 0 | 3 |

Ko iz napovedi modela vzamemo $\hat{y} = \max_i f_i(x)$, je zgornja enačba, torej napaka, enaka $1 - \hat{y}$ v primeru pravilne klasifikacije in $\hat{y}$ v primeru napačne klasifikacije. Pričakovana vrednost takšne funkcije napake je

$$E[e(\vec{x})] = p_y(\vec{x})(1 - f_y(\vec{x})) + (1 - p_y(\vec{x}))f_y(\vec{x})\,. \tag{4.2}$$

Ker so prave pogojne verjetnosti neznane za nove primere, na katerih želimo izračunati ocene zanesljivosti, nam za referenčno oceno ne preostane drugega, kot da izvedemo niz substitucij. Najboljši razpoložljivi približek prave verjetnosti razreda $p_y(\vec{x})$ je verjetnost modela $f_y(\vec{x})$. Ker pa pravi razred $y$ ni znan, se moramo zadovoljiti z napovedjo modela

$\hat{y}$. Z dvema korakoma substitucij lahko zapišemo oceno pričakovane napake:

$$\begin{aligned}\tilde{E}[e(\vec{x})] &= f_y(\vec{x})(1 - f_y(\vec{x})) + (1 - f_y(\vec{x}))f_y(\vec{x}) = \\ &= \hat{y}(1 - \hat{y}) + (1 - \hat{y})\hat{y} = \\ &= 2(\hat{y} - \hat{y}^2)\,. \end{aligned} \tag{4.3}$$

Ta referenčna ocena [52] ima dve zaželeni lastnosti. Prva je, da jo lahko izračunamo za poljubno klasifikacijsko napoved, saj je napoved modela $\hat{y}$ vedno na voljo. Druga dobra lastnost je, da je za optimalni model (ali najbližji optimalnosti) ta ocena optimalna, saj postane enaka funkciji napake. Referenčna ocena doseže svoj minimum, ko je $y$ najpreprostejša binarna spremenljivka z maksimalno entropijo (P=0.5).

Uporaba referenčne ocene je analogna uporabi relativne frekvence večinskega razreda kot reference za vrednotenje klasifikacijske uspešnosti poljubnega klasifikatorja. Na primer 90% točnost zveni dobro, vendar v primeru, da relativni delež večinskega razreda presega 90%, rezultat ni več najboljši. Podobno gre pri ocenjevanju zanesljivosti, saj če določena metoda ocenjevanja zanesljivosti ne presega rezultatov referenčne ocene, nanjo ne moremo gledati kot na uporabno oziroma koristno metodo.

## 4.3 Metodologija testiranja

V tem poglavju je testiranje izvedeno s postopkom prečnega preverjanja po metodi *izpusti-enega*. To pomeni, da vsak primer učne množice enkrat izpustimo, s preostalimi primeri pa zgradimo model, na katerem izračunamo napoved, njeno napako in vse točkovne ocene zanesljivosti izpuščenega primera. Uspešnost ocen zanesljivosti merimo s Spearman-ovim rangirnim korelacijskim koeficientom $\rho$ med ocenami zanesljivosti in klasično napako klasifikacijskih napovedi (1 - napoved pravega razreda). Spearmanov rangirni korelacijski koeficient je v bistvu Pearsonov korelacijski koeficient, izračunan nad rangiranimi podatki in je definiran kot kovarianca rangirnih spremenljivk $x$ in $y$, deljena s produktom njunih standardnih deviacij:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Na vsakem koraku prečnega preverjanja poleg ocen zanesljivosti izračunamo tudi referenčno oceno, predstavljeno v prejšnjem razdelku. Kasneje je v razdelku 4.6 analizirana korelacija med točkovnimi ocenami zanesljivosti in referenčno oceno. Če kateri

izmed korelacijskih koeficientov ni značilen, primer šteje za neznačilnega. Če sta oba koeficienta značilna, potem nadalje preverjamo njihovo razliko s klasičnim Z-testom. Pri tem testu načeloma merimo razdaljo med dvema normalnima distribucijama, kar pa zares drži le pri slabih metodah. Vendar ker želimo zgolj izločiti primere, ki so lahko produkt čistega naključja, je Z-test kljub kršenim predpostavkam primeren. Pri testu moramo med drugim upoštevati varianco podatkov, ker pa so učne množice dovolj velike (najmanjša vključuje 32 primerov), smemo uporabiti varianco vzorca.

Če Z-test zavrne, da sta korelacijska koeficienta iz iste distribucije, sta torej značilno različna, potem še preverimo, ali je korelacija ocen zanesljivosti z napako večja (boljše) ali manjša (slabše) od korelacije referenčne ocene z napako. Izbrana stopnja značilnosti je 95% ($\alpha$=0.05) tako za rangirne korelacijske koeficiente kot pri Z-testih.

Metodologija je zasnovana po vzoru [53], kjer je obravnavana primerjava več modelov na več domenah. Tukaj pa izvajamo primerjavo z referenčno oceno in preverjamo značilnost razlike s posameznimi drugimi pristopi.

Na tem mestu lahko posebej izpostavimo, da pri poskusih obravnavamo modele kot črne škatle, pa naj bodo modeli dobri ali slabi. Pri poskusih ne posvečamo nikakršne pozornosti optimizaciji posameznih modelov – modeli so venomer enaki in njihovi parametri so nastavljeni na povprečne, oz. privzete vrednosti. Zaradi tega lahko verjamemo, da so napake vseh modelov preko vseh domen naključne in eksponentno porazdeljene.

## *4.4 Vpliv različnih mer razdalj*

Lahko se porodi vprašanje, ali je pomembno, katera mera razdalje se uporablja pri enačbah 3.1, 3.3 in metodah obratne transdukcije. Kratek odgovor je, da niti ne. Daljši se pa začne z dejstvom, da je odgovor odvisen od predznanja. Seveda če že vnaprej vemo, katera mera je primerna, takrat ni razloga, da se je ne uporabi. Če to ni jasno, pa je najbolje preizkusiti vse razumne hipoteze.

Ko imamo opravka z regresijskimi problemi, ni prevelike izbire; skoraj vedno se uporablja evklidsko razdaljo, če predznanje ne pravi drugače (na primer, da bi bila Manhattan razdalja bolj robustna). Pri klasifikaciji so na voljo pestre družine mer razdalj. Cela kopica različnih mer in metrik nad verjetnostnimi distribucijami je že bila testirana na podoben način v [49].

Tu se omejimo zgolj na najpogostejše razdalje nad dvema verjetnostnima vektorjema $P$ in $Q$:

- Manhattan razdalja (prva norma)

$$\|(P,Q)\|_{man} = \sum_i |p_i - q_i|$$

- evklidska razdalja (druga norma)

$$\|(P,Q)\|_{euc} = \sqrt{\sum_i (p_i - q_i)^2}$$

- maksimalna razdalja (neskončna norma)

$$\|(P,Q)\|_{max} = max_i |p_i - q_i|$$

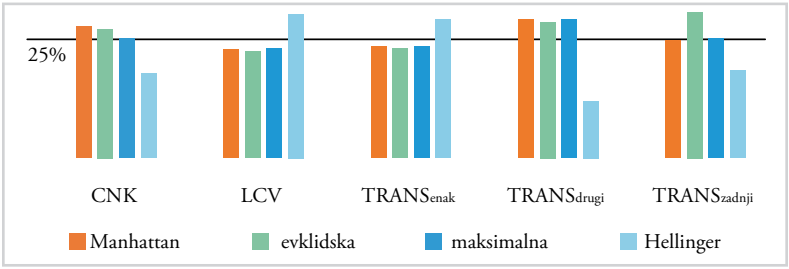
- Hellingerjeva razdalja

$$\|(P,Q)\|_{hel} = \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$$

Slika 4.1 prikazuje kakšni so doprinosi različnih mer razdalj k uspešnim rezultatom nekaterih točkovnih ocen. Iz stolpcev lahko razberemo relativno distribucijo primerov pozitivne korelacije točkovnih ocen ob uporabi različnih razdalj. Rezultati zaobjemajo napovedi vseh učnih modelov in vseh testnih množic. Iz slike lahko razberemo, da med prvimi tremi razdaljami ni bistvenih odstopanj oziroma razlik. Le pri uporabi Helingerjeve razdalje je tako, da je včasih ustrezna, včasih pa manj ustrezna. Med prvimi tremi so razlike neznačilne, zato lahko tudi pri klasifikaciji v splošnem uporabljamo standardno evklidsko razdaljo.

## 4.5 Prikazi tipičnih primerov

V splošnem je uspešnost učenja odvisna od problema oz. podatkov in samega modela oz. njegovih parametrov. Če so kršene predpostavke modela, ne moremo pričakovati dobrih rezultatov. Lahko se pa tudi zgodi, da podatki ne ustrezajo predpostavkam metod ocenjevanja zanesljivosti posameznih napovedi. Tu predstavljamo tri tipične primere.

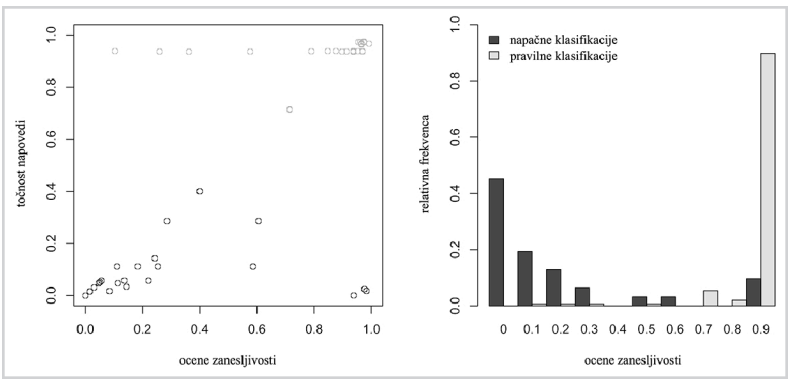



*Slika 4.1*

Relativne distribucije pozitivnih rezultatov posameznih mer razdalj.

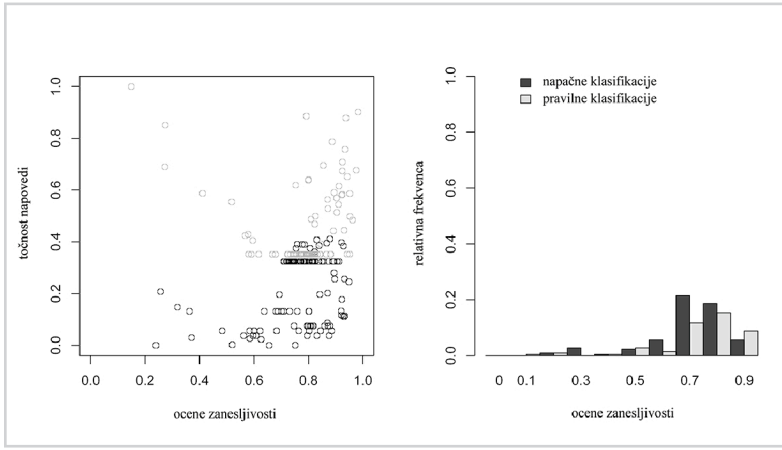
Slika 4.2 prikazuje dobri primer ob uporabi umetne nevronske mreže in metode lokalnega modeliranja napake (razdelek 3.1) z uporabo prve norme na podatkovni množici *wine*. Na levem grafu vidimo določeno stopnjo linearne korelacije med ocenami zanesljivosti in točnostjo napovedi. Na desnem grafu vidimo, da je večina napačnih napovedi na levi in pravilne napovedi na desni. To pomeni, da te ocene zanesljivosti dobro ločijo pravilne in nepravilne napovedi.

Pri veliki večini poskusov rezultati niso tako dobri. Le redko ocene zanesljivosti bolje napovejo točnost napovedi, kot uporabljeni modeli sami. Primer takšnega negativnega primera je prikazan na sliki 4.3, ponovno z uporabo umetne nevronske mreže in metode lokalnega modeliranja napake z uporabo druge norme, vendar na podatkovni množici *glass*. Za razliko od prejšnje slike tu ne vidimo nikakršne korelacije med ocenami zanesljivosti in točnostjo napovedi. Na desnem grafu je razvidno, da te ocene




*Slika 4.2*

Primer značilno pozitivne korelacije med ocenami zanesljivosti in točnostjo napovedi ter značilno dobre ločljivosti pravilnih napovedi od nepravilnih ($\rho = 0.86$; uporabljena je umetna nevronska mreža, lokalno ocenjevanje napake s prvo normo in množica *wine*).




*Slika 4.3*

Primer neznačilne korelacije med ocenami zanesljivosti in točnostjo napovedi ter slabe ločljivosti pravilnih napovedi od nepravilnih ($\rho = 0.29$; uporabljena je umetna nevronska mreža, lokalno ocenjevanje napake z drugo normo in množica *glass*).

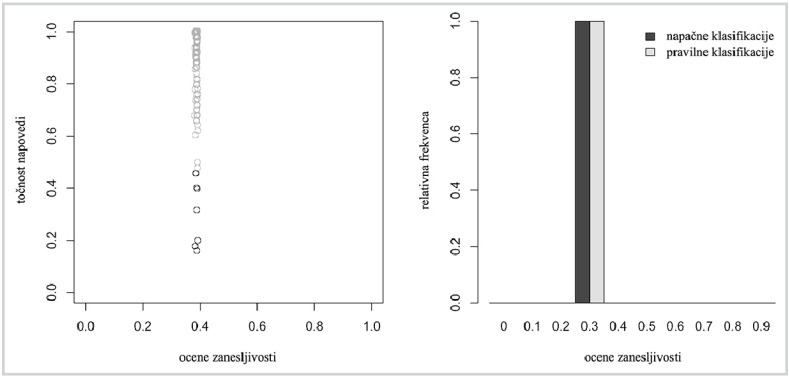
zanesljivosti niso zmožne ločiti med pravilnimi in nepravilnimi napovedmi.

Pri nekaterih poskusih pa so rezultati še slabši. Slika 4.4 prikazuje obnašanje metode ocenjevanja zanesljivosti na osnovi gostote učnih primerov na že videni množici *wine*. Ker metoda sloni na predpostavki, da je zanesljivost napovedi odvisna od različnih gostot učnih primerov, je metoda popolnoma neuporabna, ko so učni primeri enakomerno porazdeljeni. Ko je učna množica uniformno gosta, so vse ocene zanesljivosti enake in posledično kakršnakoli ločljivost pravilnih in nepravilnih napovedi ni možna.

## 4.6 Uporabnost točkovnih cenilk

V razdelku 4.3 smo omenili porazdelitve rezultatov ki jih analiziramo. Pristopi, ki ne nudijo dobrih informacij o napaki posamezne napovedi, ne morejo imeti dobre korelacije z napako in njihovi korelacijski koeficienti so zato normalno porazdeljeni okoli 0. Boljša kot je metoda, več informacij nudi o napaki in povprečna korelacija je višja.

Slika 4.5 prikazuje stolpčne diagrame, ki vsebujejo za vsak posamezni pristop dosežene korelacijske koeficiente med ocenami in napako, za vse modele in vse domene. Očitno je, da so korelacijski koeficienti pristopa na osnovi gostote normalno porazdeljeni okoli 0, kar pomeni da so doseženi korelacijski koeficienti produkt naključja.




*Slika 4.4*

Primer ko metoda ocenjevanja zanesljivosti na osnovi gostote učne množice ne deluje (uporabljen je model bagging na množici *wine*).

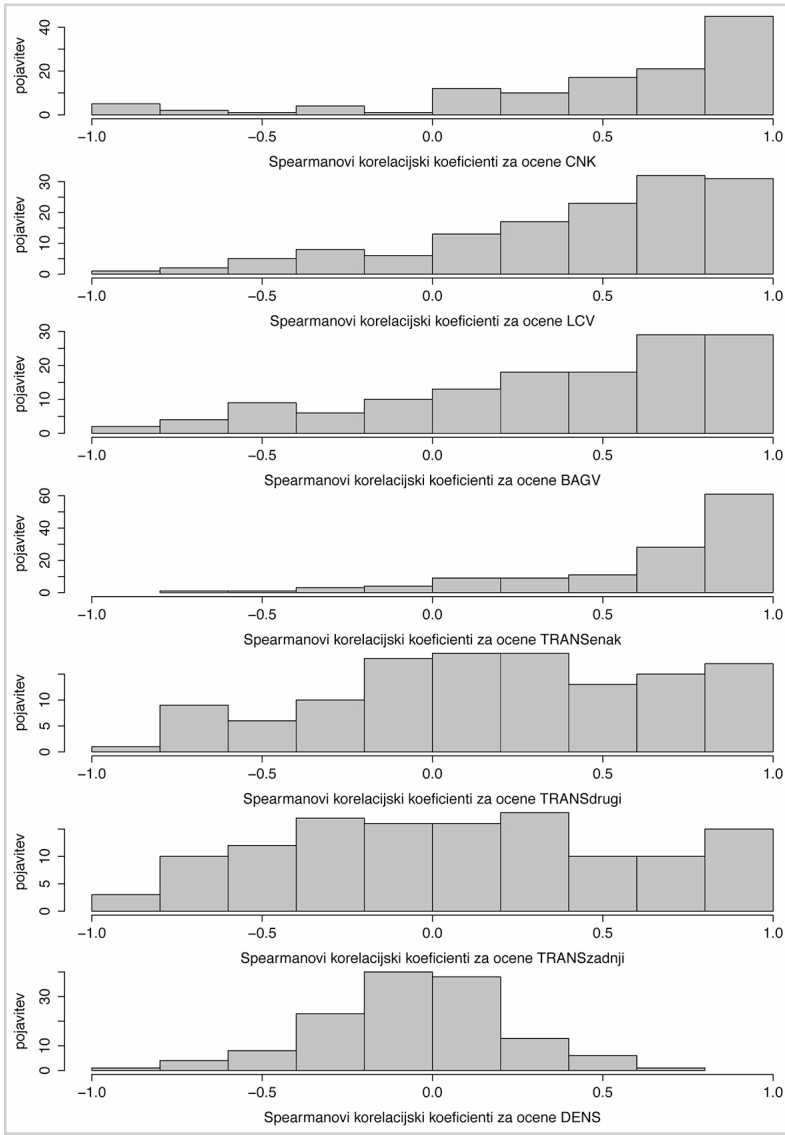
Tudi za pristopa $TRANS_{drugi}$ in $TRANS_{zadnji}$ je razvidno, da del poskusov prihaja iz normalne distribucije okoli 0, vendar pa vseeno vidimo, da je drugi del distribucije nagnjen proti 1. Originalni pristop $TRANS_{enak}$ pa ima izmed vseh metod najbolj ugodno distribucijo korelacije z napako.

Glavno vprašanje tega poglavja je, ali so točkovne ocene zanesljivosti primerne oziroma dobre kot indikatorji zanesljivosti posameznih napovedi. V razdelku 4.3 je bil predstavljen postopek, po katerem so poskusi izpeljani z vsemi modeli, metodami ocenjevanja zanesljivosti, merami razdalj in podatkovnimi množicami. Ekvivalentna analiza, vendar opravljena le na modelu umetnih nevronskih mrež, je predstavljena v [52].

Tabela 4.2 prikazuje glavne rezultate in beremo jo na sledeč način: ocene zanesljivosti po metodi lokalnega modeliranja napake napovedi so v 8% poskusov dosegle značilno višjo korelacijo med ocenami zanesljivost in točnostjo napovedi, kot jo je dosegla referenčna ocena. V 57% primerov je cenilka dosegla značilno nižjo korelacijo, kot jo je dosegla referenčna ocena. V 6% primerov sta obe korelaciji sicer značilni, vendar po Z-testu nista značilno različni (zato oznaka *enako*). V zadnjem stolpcu so zabeleženi odstotki primerov, v katerih vsaj eden izmed korelacijskih koeficientov ni značilen in zato nadaljnja primerjava ni smotrna.

Zadnji dve vrstici ne prikazujeta rezultatov metod temveč napovedi samih modelov in referenčne ocene. Vidimo, da v 9% poskusov referenčna ocena ni dosegla značilne korelacije s točnostjo napovedi, kar postavi zgornjo mejo uspešnosti cenilk zanesljivosti na 91%. Rezultati so še grafično prikazani na sliki 4.6.

Rezultati lahko komu nakazujejo, da točkovne cenilke niso uspešni indikatorji za-




*Slika 4.5*

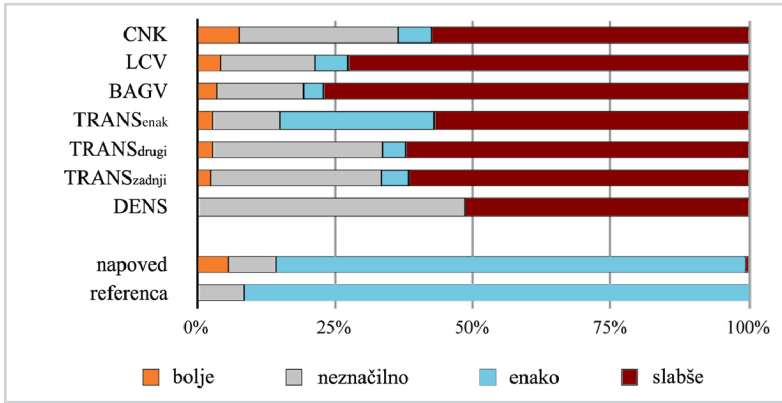
Histogrami korelacijskih koeficientov med ocenami in napako.

nesljivosti posameznih napovedi. Potrebno je izpostaviti, da uporabljamo tak nabor modelov, kjer želimo tudi neprimerne in ne optimalne napovedi. Zabeležili smo sicer nizek odstotek, pri katerem so točkovne ocene izboljšale napovedi, vendar ti rezultati bolj opozarjajo na pomembnost obravnavanja vsake problemske množice posebej ter analize posameznih kombinacij modelov in njihovih parametrizacij.

*Tabela 4.2*

Eksperimentalna evalvacija točkovnih ocen zanesljivosti. Korelacijo ocen zanesljivosti in točnosti napovedi primerjamo s korelacijami referenčne ocene. Števila predstavljajo odstotke, vsaka vrstica se sešteje v 100%.

| | bolje | slabše | enako | neznačilno |
|---|---|---|---|---|
| CNK | 8 | 57 | 6 | 29 |
| LCV | 4 | 73 | 4 | 17 |
| BAGV | 4 | 76 | 4 | 16 |
| TRANS prvi | 3 | 57 | 28 | 12 |
| TRANS drugi | 3 | 62 | 4 | 31 |
| TRANS tretji | 2 | 62 | 5 | 31 |
| DENS | 0 | 51 | 0 | 49 |
| napoved | 6 | 1 | 85 | 8 |
| referenčna ocena | - | - | 91 | 9 |




*Slika 4.6*

Vizualizacija evalvacije točkovnih cenilk. Korelacijo ocen zanesljivosti in točnosti napovedi primerjamo s korelacijami referenčne ocene. Vsi eksperimenti predstavljajo 100%, prikazani so relativni deleži pozitivnih, neznačilnih, enakih in negativnih rezultatov.

5

# *Intervalne cenilke zanesljivosti*

V tem in naslednjih dveh poglavjih se omejimo na regresijske probleme, saj nad diskretnimi vrednostmi intervali niso naravno definirani. Zato se osredotočamo na splošni regresijski primer, kjer imamo opravka s skrito funkcijo $f(\vec{x})$, ki slika atributni prostor $\vec{x}$ na realno spremenljivko $y$. Z vzorčenjem te funkcije dobimo množico vrednosti $(\vec{x}_i, y_i), i = 1..n$, ki pa so lahko šumna opazovanja:

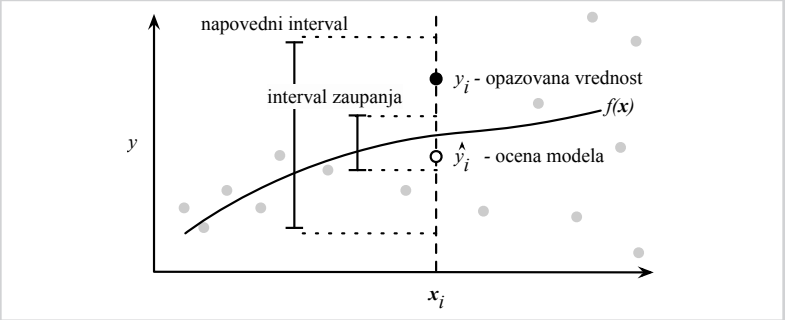
$$y_i = f(\vec{x}_i) + \varepsilon(\vec{x}_i)\,.$$

S stališča zanesljivosti posameznih napovedi so za nas zanimivi tako imenovani *napovedni intervali* (angl. *prediction intervals*, zato včasih tudi *predikcijski intervali*), katerih naloga je opisati interval vrednosti, znotraj katerega bi se z verjetnostjo $1 - \alpha\%$ (tradicionalno 90%, 95% in 99%) morala znajti prava vrednost novega, neznanega primera.

Če za šum predpostavimo, da prihaja iz Gaussove distribucije, torej da je njegova pričakovana vrednost 0 in da je varianca konstantna v problemskem prostoru, potem imamo opravka z lahkim problemom. Na tej preprosti predpostavki temelji najbolj preprost regresijski model, linearna regresija, pa tudi večina drugih modelov. Ob tej predpostavki je napovedne intervale sila preprosto postaviti in odgovor najdemo v poljubnem statističnem priročniku, oziroma ga razberemo iz tabele standardnih vrednosti (Z-vrednosti).

Večina objavljene literature sloni na tej predpostavki *homoskedastičnosti* podatkov. Če ne želimo narediti tako močne predpostavke in dopuščamo možnost, da šum le ni tako preprost, potem so zadeve bolj zapletene. Kljub temu, da nepristranski šum ne vpliva bistveno na ocenjevanje pogojne pričakovane vrednosti in posledično na točnost napovedi, kakršne koli klasične ocene zanesljivosti (ki temeljijo na homoskedastičnosti) zavajajo uporabnike in jim oblikujejo napačne predstave.

Šum v *heteroskedastičnih* podatkih lahko razbijemo na dve neodvisni komponenti, kjer vsaka vnaša določeno stopnjo šuma. Prvo komponento imenujemo *varianca negotovosti modela* in označujemo s $\sigma_m^2$. Tu imamo v mislih faktorje, ki vplivajo na model. Izvori tega šuma so neregularnost podatkov za določen model, pristranskost modela, njegova odpornost na šum, prilagajanje šumu, zmožnost posploševanja in podobno. Drugo komponento imenujemo *varianca šuma podatkov*, označujemo pa s $\sigma_p^2$. Na njo vplivajo različni zunanji dejavniki, povzročajo pa različne stopnje točnosti v različnih predelih problemskega prostora. Ker predpostavljamo neodvisnost teh dveh komponent, ju lahko preprosto seštejemo in imamo opravka s *skupno varianco napovedi*: $\sigma^2 = \sigma_m^2 + \sigma_d^2$.




*Slika 5.1*

Ilustracija intervalov zaupanja in napovednih intervalov.

Slika 5.1 prikazuje model, po katerem se ravnamo. Prikazani so intervali zaupanja, napovedni intervali in tri ključne vrednosti. Intervali zaupanja obravnavajo točnost modelovih ocen $\hat{y}_i = \hat{y}(\vec{x}_i)$ prave, vendar skrite funkcije $f(\vec{x}_i)$. Zatorej intervali zaupanja skušajo čim bolje opisati varianco modela, oziroma distribucijo vrednosti $f(\vec{x}_i) - \hat{y}_i$. V praktičnih primerih pa je bolj informativna in koristna primerjava točnosti modelovih napovedi z dejanskimi opazovanimi vrednostmi $y_i$. Napovedni intervali se trudijo zajeti distribucijo vrednosti bodočih posameznih primerov, oziroma njihovih odstopanj od modelovih napovedi; torej opisujejo distribucijo vrednosti $y_i - \hat{y}_i$. Če razširimo prvi člen,

$$y_i - \hat{y}_i = f(\vec{x}_i) + \varepsilon(\vec{x}) - \hat{y}_i = \left[f(\vec{x}_i) - \hat{y}_i\right] + \varepsilon(\vec{x}),$$

vidimo, da je ustrezni interval zaupanja zajet znotraj intervala napovedi. S stališča praktične uporabe so napovedni intervali bolj informativni, saj opisujejo pričakovano distribucijo bodočih, neznanih vrednosti, za razliko od intervalov zaupanja, katerih skrb je natančnost ocenjevanja pogojnih povprečnih vrednosti.

Pri ocenah časovnih zahtevnosti ponovno uporabljamo $\mathcal{O}$ notacijo, kjer $n$ predstavlja število razpoložljivih učnih primerov, $m$ število atributov, za model $M$ pa z $\mathcal{O}(M)$ označujemo časovno zahtevnost napovednega algoritma in z $\mathcal{O}(M_n)$ časovno zahtevnost učnega algoritma.

## 5.1 Stremljenje in maksimalno verjetje

Zgodovinsko gledano so se raziskovalci najprej lotili ocenjevanja vsake komponente skupne variance napovedi posebej. Dobro oceno variance negotovosti modela, $\sigma_m^2(\vec{x})$,

je moč dobiti z metodami stremljenja (angl. *bootstraping*), oceno variance šuma podatkov, $\sigma_p^2(\vec{x})$, pa z metodami ocenjevanja maksimalnega verjetja (angl. *maximum-likelihood estimation*). V naslednjih razdelkih sta predstavljeni dve različici takšnega pristopa k tvorbi intervalnih ocen zanesljivosti.

### 5.1.1 Poenostavljena metoda (SMVa)

Najbolj preprosta različica uporabe stremljenja in ocenjevanja maksimalnega verjetja je predstavljena v [39]. Za dani model, katerega učni algoritem nam je na voljo, in za podano učno množico, je posplošena reinterpretacija njihove metode sledeča.

Najprej se posvetimo varianci negotovosti modela. Za cenilko vrednosti $\sigma_m^2(\vec{x})$ na učnih podatkih izvedemo bagging s pomočjo danega učnega algoritma, kar tvori napoved $\hat{y}_{bag}(\vec{x})$. Intervale zaupanja se tvori tako, da upoštevamo normalno distribucijo posameznih napovedi znotraj modela bagging. Varianca teh napovedi predstavlja varianco negotovosti modela. Predpostavka normalnosti je pri nesimetrični distribuciji residualov kršena, kar pa se izrazi v širših intervalih.

Iz napovedi modela bagging pridobimo množico residualov (razlike napovedi od pravih vrednosti): $y(\vec{x}) - \hat{y}_{\text{bag}}(\vec{x})$ vseh primerov učne množice. Ti residuali nosijo informacijo o varianci šuma podatkov, zato se za čim boljšo oceno te vrednosti uporablja pristop ocenjevanja maksimalnega verjetja. Omenjeni avtorji so za ta korak uporabili umetne nevronske mreže (od koder izvirajo začetna dela s področja napovednih intervalov), ki so oblika ocenjevanja maksimalnega verjetja. Ker lahko uporabimo poljuben pristop ocenjevanja maksimalnega verjetja, za ocenjevanje $\sigma_p^2$ raje uporabljamo mrežo radialnih baznih funkcij (iz razdelka 2.2.3), ki je nekoliko bolj robusten in bolj splošen model, saj uporablja manj parametrov.

Predpostavili smo, da sta komponenti neodvisni, zato je lokalna ocena variance $\hat{\sigma}^2(\vec{x}) = \hat{\sigma}_m^2(\vec{x}) + \hat{\sigma}_p^2(\vec{x})$. Za oblikovanje končnih napovednih intervalov privzamemo, da residuali in negotovost modela z večanjem števila upoštevanih primerov konvergirajo po centralnem limitnem izreku k lokalno normalno porazdeljeni distribuciji. Tako je napovedni interval (označujemo ga z *SMVa*) definiran kot

$$\begin{aligned} &\hat{y}_{\text{bag}}(\vec{x}) \pm z_{\frac{\alpha}{2}}\hat{\sigma}^2(\vec{x}) = \\ &\hat{y}_{\text{bag}}(\vec{x}) \pm z_{\frac{\alpha}{2}}(\hat{\sigma}_m^2(\vec{x}) + \hat{\sigma}_p^2(\vec{x})), \end{aligned} \tag{5.1}$$

kjer je $z$ funkcija kumulativne porazdelitve normalne distribucije.

Kar se časovne zahtevnosti tiče, je stremljenje precej drago, saj za prvi korak ($b = 30$ vzorčenj) potrebujemo čas $\mathcal{O}(b \cdot (M_n + n \cdot M))$ in dobimo $b \cdot n$ residualov. Ti residuali tvorijo bazne vektorje mreže radialnih baznih funkcij, za njihov raztros pa moramo izračunati razdalje med vsemi residuali, za kar potrebujemo čas reda največ $\mathcal{O}(b^2 n^2)$. Ko to opravimo, je izračun napovednega intervala manj zahteven, saj potrebuje le napoved in mrežo radialnih baznih funkcij; potrebuje čas reda $\mathcal{O}(M + b \cdot n + 1)$.

### 5.1.2 *Originalna metoda (SMVb)*

Zgodovinsko gledano je bila prva in nekoliko bolj zapletena metoda objavljena v [38]. Poenostavljena metoda se od originalne loči v dveh bistvenih elementih.

Cenilka negotovosti modela je enaka; oceno $\sigma_m^2(\vec{x})$ tvori varianca notranjih napovedi modela bagging. Prva bistvena razlika je pri ocenjevanju variance šuma podatkov. Kjer smo pri poenostavljeni metodi množico residualov pridobili iz vseh učnih primerov z napovedmi modela bagging, se tu izkorišča dejstvo, da vzorčenje s ponavljanjem ne uporabi čisto vseh primerov učne množice. Ti primeri niso uporabljeni za učenje modelov, zato njihove napovedi nudijo boljšo, nepristransko množico residualov. Za ocenjevanje $\sigma_p^2$ tudi tu uporabljamo mrežo radialnih baznih funkcij, vendar z nepristranskimi residuali.

Druga razlika je v tem, kje je postavljena sredina intervalov. Pri prejšnji metodi je bilo predpostavljeno, da $\hat{y}_{\text{bag}}$ nudi bolj stabilno oceno prave funkcije $f(\vec{x})$. Originalna metoda pa postavlja napoved modela v sredino in napovedni intervali, poimenovani *SMVb*, so tako

$$\begin{aligned} &\hat{y}(\vec{x}) \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\vec{x}) = \\ &\hat{y}(\vec{x}) \pm z_{\frac{\alpha}{2}} (\hat{\sigma}_m^2(\vec{x}) + \hat{\sigma}_p^2(\vec{x})), \end{aligned} \tag{5.2}$$

kjer sta $\hat{\sigma}^2(\vec{x}) = \hat{\sigma}_m^2(\vec{x}) + \hat{\sigma}_p^2(\vec{x})$ in $z_{\frac{\alpha}{2}}$ že prej omenjena standardna vrednost za izbrano $\alpha$.

Analiza časovne zahtevnosti je enaka kot za poenostavljeno metodo, vendar imamo zaradi uporabe primerov, ki niso v vzorcih, opravka z manj residuali (pridobimo konstantni faktor $\frac{1}{3}$).

## 5.2 Lokalne okolice

Metode na osnovi stremljenja so se izkazale za dobre, vendar precej časovno zahtevne. V realnem svetu pogosto stremimo k časovno bolj učinkovitim metodam, zato se v tem razdelku posvečamo preprostejšim metodam. Te se osredotočajo na neposredno ocenjevanje skupne variance napovedi in temeljijo na lokalnih okolicah. Zaradi svoje časovne učinkovitosti jih je možno uporabljati tudi takrat, ko je stremljenje časovno nesprejemljivo.

### 5.2.1 Najbližji sosedi (NS5 in NS100)

Metoda temelji na preprosti predpostavki, da najbližji sosedi lahko ponudijo informacijo o lokalni zanesljivosti napovedi posameznega primera, oziroma da se bo prava vrednost novega primera iz iste distribucije kot za že znane bližnje primere. Napovedne intervale tvorimo po sledečem postopku.

Najprej izračunamo predznačene residuale celotne učne množice. Povprečna vrednost residualov bližnjih primerov, $\bar{r}$, služi za odpravljanje pristranskosti in z njo popravljamo sredino napovednega intervala. Za residuale predpostavimo da so normalno porazdeljeni, zato nam njihova varianca služi kot ocena $\hat{\sigma}^2(\vec{x})$. Slednja vrednost definira širino napovednega intervala, katerega lahko zdaj zapišemo kot

$$\hat{y}(\vec{x}) + \bar{r} \pm z_{\frac{\alpha}{2}} \hat{\sigma}^2(\vec{x}). \tag{5.3}$$

Velikost izbrane okolice je ključni parameter metode, uporabimo pa enak pristop, kot v razdelku 3.2. Pri testiranjih smo uporabili dve vrednosti. Metoda z oznako *NS5* uporablja 5% podatkov, ki so na razpolago. Druga inačica, *NS100*, tvori napovedne intervale z uporabo celotne učne množice. Kot takšna je ekvivalentna preprostim analitičnim pristopom, ki predpostavljajo konstantno varianco, zaradi tega pa je računsko še manj zahtevna. Časovna zahtevnost metode je pogojena z iskanjem najbližjih sosedov, torej reda $\mathcal{O}(n \cdot (m \log n) \cdot 1 + 1) = \mathcal{O}(nm \log n)$.

Omeniti še moramo, da za pristop ni nujno, da napovedni intervali vključujejo napoved modela. Intervali se prilagajajo izmerjenim vrednostim odvisne spremenljivke in korekcija pristranskosti lahko povzroči, da sama napoved modela pristane izven napovednega intervala (in je po vsej verjetnosti slaba).

### 5.2.2 Razvrščanje v skupine (NSrs)

Ta pristop postavljanja napovedi, predstavljen v [40], je bil inspiracija za metodo, predstavljeno v prejšnjem razdelku. Tu uporabljamo idejo razvrščanja v skupine, vendar za razliko od izvirne metode, ki uporablja mehke množice, tu uporabljamo preprostejše, klasično razvrščanje v skupine (s pomočjo $k$ centroidov). V prid preprostosti metode ne vpeljujemo nobene optimizacije razvrščanja. Problemu, oziroma vprašanju optimalnega števila skupin se izognemo z uporabo splošne hevristike, po kateri podatke razvrstimo v $k = \sqrt{\frac{n}{2}}$ skupin, kjer je $n$ velikost učne množice.

Napovedne intervale *NSrs* se tvori za vsako skupino posebej, direktno iz empirične distribucije residualov. Če želimo na primer 90% napovedne intervale, je $\alpha = 0.1$ in zato napovedni interval definirata 5. in 95. percentil skupine residualov. Napovedni interval novega primera definira skupina, v katero je ta primer razvrščen.

Časovna zahtevnost razvrščanja v skupine [54] je $\mathcal{O}(\sqrt{\frac{n}{2}}^2 nm\sqrt{\frac{n}{2}}) = \mathcal{O}(n^{2\frac{1}{2}}m)$. Nato je potrebno vse residuale znotraj skupin urediti, kar terja čas reda $\mathcal{O}(\sqrt{\frac{n}{2}} \cdot \sqrt{2n} \log \sqrt{2n})$. Za napovedni interval novega primera moramo poiskati pripadajočo skupino, za kar je potreben čas reda $\mathcal{O}(m \log \sqrt{\frac{n}{2}})$.

### 5.2.3 Kvantilni regresijski gozdovi (KRG)

Ogledali smo si že dva pristopa k intervalnemu ocenjevanju zanesljivosti na podlagi lokalnih okolic. Tretji korak je lahko uporaba adaptivnih okolic. Vsaka adaptacija, prilagodljivost okolic, je na nek način parametrizirana, vendar v [43] najdemo reinterpretacijo naključnih gozdov, ki temelji na uteženem povprečju opazovane zvezne spremenljivke.

Kvantilni regresijski gozdovi (angleško *Quantile Regression Forest—KRG*) so bili predstavljeni v delu [44]. Za razliko od modela naključnega gozda (iz razdelka 2.2.3), listi dreves tega modela ohranijo vsa opazovanja (vse učne primerke). S tem se ohrani več informacije o skriti distribuciji in omogoči ocenjevanje *pogojnih kvantilov*. Uporabljamo jih za učenje residualov (označeni z $r$ oz. $R$), saj lahko funkcijo njihove porazdelitve pogojno z $X = \vec{x}$ zapišemo kot

$$F(r|X = \vec{x}) = P(R \leq r|X = \vec{\mathbf{x}}) = E(1_{\{R \leq r\}}|X = \vec{x}). \tag{5.4}$$

Izraz 5.4 lahko aproksimiramo z uteženim povprečjem vrednosti zabeleženih indikatorskih spremenljivk, s cenilko

$$\hat{F}(r|X = \vec{x}) = \sum_{i=1}^{b} w_i(\vec{x}) \cdot 1_{\{R_i \leq r\}}, \tag{5.5}$$

kjer $1_{\{R_i \leq r\}}$ predstavlja indikatorske spremenljivke

$$1_{\{R_i \leq r\}} = \begin{cases} 1 & R_i \leq r, \\ 0 & \text{sicer.} \end{cases}$$

Uteži $w_i$ so uteži posameznih dreves, pri čemer je utež posameznega drevesa relativna frekvenca števila primerov v listu. Vsota vseh uteži drevesa je ena, tako da skupaj tvorijo verjetnostni vektor oziroma distribucijo residualov. Oceno pogojnih kvantilov dobimo tako, da poiščemo infimalno vrednost $r$, za katero velja $\hat{F}(r|X = \vec{x}) \geq \alpha$.

Uteži pridobimo tako, da novi primer spustimo po drevesih gozda. Vse $r_i$ v listih najprej uredimo po naraščajočih vrednostih. Ko tvorimo $1 - \alpha$ napovedne intervale, moramo poiskati vrednosti $r_s$ in $r_z$, za kateri velja $\sum_{i=1}^{s} w_i \geq \frac{\alpha}{2}$ in $\sum_{i=1}^{z} w_i \geq 1 - \frac{\alpha}{2}$. Spodnja meja napovednega intervala je $\hat{y} + r_s$, zgornja meja pa $\hat{y} + r_z$.

Časovna zahtevnost tega algoritma formalno ni bila izpeljana, zato podajamo le oceno. Gradnja posameznega drevesa potrebuje čas reda $\mathcal{O}(nm \log n)$, iskanje v drevesu pa $\mathcal{O}(\log n)$, uporaba gozda pa doprinese konstantni faktor (privzeto $b$=1000 dreves). Ker izračun napovednega intervala traja konstanten čas, ko preiščemo vsa drevesa, je potreben čas reda $\mathcal{O}(\log n)$.

# 6

# *Osnovni poskusi na regresijskih problemih*

V tem poglavju predstavimo delovanje in obnašanje intervalnih cenilk zanesljivosti posameznih napovedi, ki so bile opisane v prejšnjem poglavju, na množici regresijskih problemov. V prvem razdelku so predstavljene umetne in realne testne množice, temu pa sledi predstavitev statistik, s katerimi vrednotimo uspešnost intervalnih ocen. Sledi demonstracija delovanja intervalnih cenilk s pomočjo vizualizacij. Poglavje zaključuje opis metodologije testiranja in osnovni rezultati, oziroma analiza doseženih mer uspešnosti. Delo, predstavljeno v tem poglavju, bo objavljeno v [55].

## 6.1 Testne množice

Za poskuse z intervalnimi cenilkami smo uporabili 156 regresijskih množic podatkov. Med njimi je 30 realnih podatkovnih množic, ki so zbrane iz repozitorija *UCI* [50] ter zbirk *numeric* in *regression* odprtokodnega projekta *Weka* [56], zanje pa je postavljen pogoj, da vsebujejo vsaj 50 primerov. Največja podatkovna množica, *concrete*, ima 1030 primerov; povprečno št. primerov je 267.

Preostalih 126 podatkovnih množic tvori šest različnih umetnih regresijskih problemov, kjer je iz vsake izmed funkcij vzorčenih 50, 100, 200, 400, 800 in 1600 primerov. Pet vrst umetnih problemov je dvorazsežnih, da jih lahko grafično analiziramo; najbolj preprosta sta vzorčena iz linearne funkcije z dodanim homogenim in heterogenim (linearno odvisnim od $x$) Gaussovim šumom. Nato imamo tri-delno odsekovno funkcijo s heterogenim šumom, odsekoma konstantno funkcijo z različnimi stopnjami Gaussovega šuma ter nelinearno, sinusoidno funkcijo s šumom, vzorčenim iz mešanice treh različnih Gaussovih distribucij. Teh pet problemov je dodatno razširjenih v različice z dodanimi enim, dvema in tremi različnimi, neodvisnimi naključnimi atributi. Za zadnjo umetno množico je zamišljeno, da je težka, tvori jo deset atributov, izmed katerih je šest vzorčenih iz različnih distribucij in so med seboj močno odvisni (kombinacije normalnih, enakomernih, beta, in Poissonovih distribucij, tudi s sinusno transformacijo); preostali štirje so podvojeni atributi, z dodanim multiplikativnim in aditivnim šumom iz različnih distribucij.

Umetni problemi se razlikujejo po kompleksnosti in stopnjah šuma. S poznavanjem delovanja posameznih modelov je možno imeti neko predstavo, kateri modeli bodo dosegali boljše rezultate od drugih in na katerih problemih; trivialni primer je, da bo linearna regresija najbolj ustrezna za preprosto podatkovno množico linearne funkcije, na isti množici pa ne pričakujemo prav dobrih rezultatov regresijskih dreves. Tako so umetne množice primarno uporabljane za validacijo. Primeri iz realnega sveta so precej

bolj kompleksni in so uporabljeni za preverbo praktične uporabnosti intervalnih cenilk zanesljivosti.

## 6.2 Mere uspešnosti

Pri ocenjevanju uspešnosti napovednih intervalov nas zanimata dve vrednosti. Prva mera je *pokrivna verjetnost napovednih intervalov* (PVNI; angleško *Prediction Interval Coverage Probability*) in je definirana kot odstotek testnih primerov, za katere je prava vrednost odvisne spremenljivke zajeta znotraj napovednih intervalov. Iz same definicije napovednih intervalov od njih pričakujemo, da v povprečju zajamejo vnaprej določen odstotek, $(1 - \alpha)\%$, realiziranih opazovanj, zato je pokrivna verjetnost napovednih intervalov kvantizacija *pravilnosti*. Bližje kot se napovedni intervali približajo zadani pokrivni verjetnosti, bolj so pravilni.

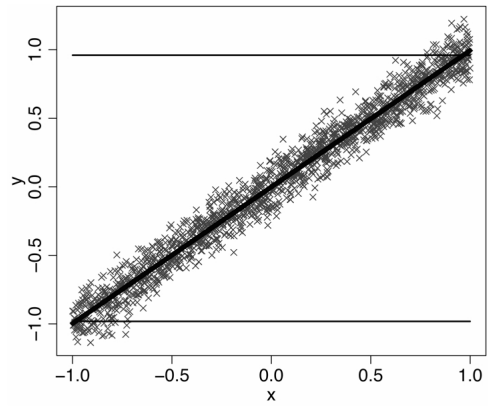
Druga mera opisuje povprečno širino napovednih intervalov in ji pravimo *povprečni napovedni interval* (angleško *Mean Prediction Interval*). Ker so ožji intervali praviloma bolj zaželeni, je ta mera kvantizacija *optimalnosti*. Ker želimo primerjati to količino med različnimi pristopi na različnih problemskih domenah, potrebujemo nek skupni imenovalec za normalizacijo. V ta namen konstruiramo *privzeti napovedni interval* neke učne množice direktno iz empirične distribucije razpoložljivih opazovanj odvisne spremenljivke, tako da vzamemo ustrezna kvantila (na primer 5. in 95. percentil za 90% interval). Povprečni napovedni interval, normaliziran s privzetim napovednim intervalom, imenujemo *relativni povprečni napovedni interval* (RPNI; angleško *Relative Mean Prediction Interval*). Manjša kot je ta vrednost, bolj so intervali določene metode optimalni.

Po kriterijih teh dveh mer uspešnosti doseže najboljši rezultat tista metoda, katere pokrivna verjetnost napovednih intervalov je najbližje zadani verjetnosti, relativni napovedni interval pa je najmanjši možen.

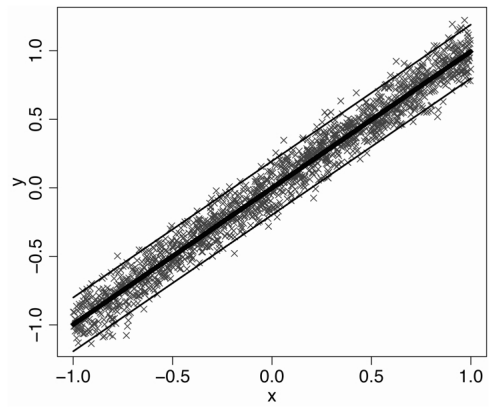
## 6.3 Prikaz z vizualizacijami

Razdelek je namenjen grafični predstavitvi intervalnih cenilk zanesljivosti, s katerimi razkrivamo delovanje in obnašanje različnih pristopov ter kako to vpliva na mere uspešnosti.

Slika 6.1 najprej prikazuje razliko med konstantnimi (katere uporabljamo za normalizacijo pri meri optimalnosti) in intervali, odvisnimi od vhodnih podatkov, na najbolj



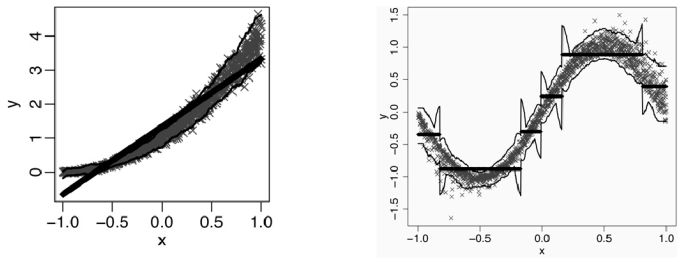

(a) PVNI=0.950, RPNI=1




(b) PVNI=0.945, RPNI=0.197

*Slika 6.1*

Prikaz razlike med konstantnimi napovednimi intervali in napovednimi intervali, odvisnimi od vhodnih podatkov, na linearnem problemu $y = x + \mathcal{N}(0, 0.1)$ in z linearnim modelom. Opazovane vrednosti so sive točke, napovedi modela so črne točke in črte so napovedni intervali. Sliki prikazujeta (a) privzeti napovedni interval in (b) napovedni intervali metode NS100.




(a) PVNI=0.958 (b) PVNI=0.953

*Slika 6.2*

Prikaz napovednih intervalov po metodi NS5. Na slikah so (a) linearna regresija na problemu kvadratne funkcije z linearno odvisnim šumom in (b) regresijsko drevo na trigonometričnem problemu z mešanim šumom.

preprostem problemu linearne odvisnosti z normalno porazdeljenim, homogenim šumom. Na levi sliki vidimo privzeti napovedni interval, na desni pa analitično izračunan napovedni interval na osnovi variance učne množice. Pri vseh primerih je izbran $\alpha$ enak 0.05 — prikazani so 95% napovedni intervali.

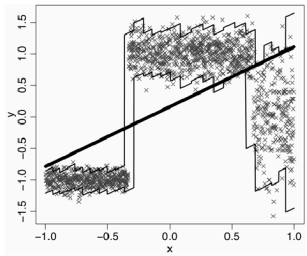
Naslednji par vizualizacij na sliki 6.2 prikazuje, kako se napovedni intervali, ki uporabljajo 5% najbližjih sosedov, prilegajo dejanskim podatkom. Iz leve slike 6.2 je razvidno, da se model linearne regresije ne nauči pravega koncepta in da so napovedi posledično napačne, oziroma ne sledijo pravi funkciji.

Napovedni intervali po metodah lokalnih okolic upoštevajo pristranskost modela, zato se na sliki vidi, kot da so intervali neodvisni od napovedi modelov. Vendar je iz desne slike 6.2 razvidno, da se pristop posveča residualom, kateri so pričakovano večji na robovih listov dreves.

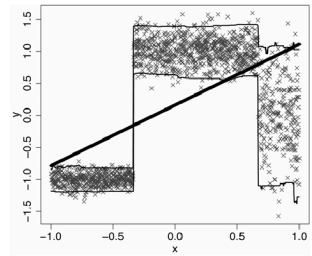
Na sliki 6.3 so na levi prikazani napovedni intervali s pristopom razvrščanja v skupine. Razvidno je, da se tudi ta pristop bolj prilagaja podatkom kakor napovedim modela. Kljub temu, da je optimalno število skupin za uporabljano odsekovno konstantno funkcijo 3, metoda dosega solidno pokrivno verjetnost, relativni povprečni napovedni interval pa znaša, 0.417, kar je blizu teoretično pravilnemu.

Na sliki 6.3 so na desni prikazani napovedni intervali kvantilnega regresijskega gozda. Vidno je, da so ti intervali bolj robustni in se bolje prilegajo residualom, zato ni presenetljivo, da so pravilnejši in optimalnejši (RPNI=0.367).

Pristopa stremljenja sta prikazana na sliki 6.4. Za oba je razvidno, da so napovedni intervali simetrični okoli napovedi. Med metodama obstaja bistvena razlika v prileganju residualom. Na levi sliki vidimo, da so napovedni intervali relativno homogeni



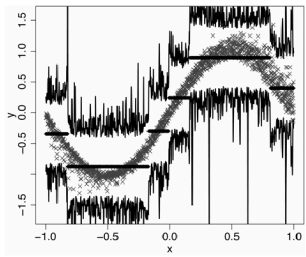

(a) PVNI=0.936



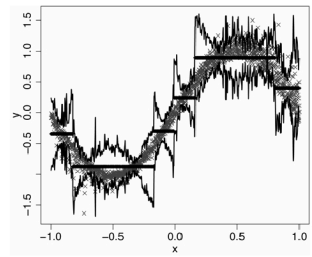

(b) PVNI=0.949

*Slika 6.3*

Napovedni intervali za linearno regresijo na odsekovni funkciji (a) pristopa z razvrščanjem v skupine in (b) kvantilnega regresijskega gozda.



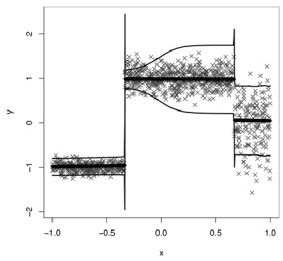

(a) PVNI=0.998



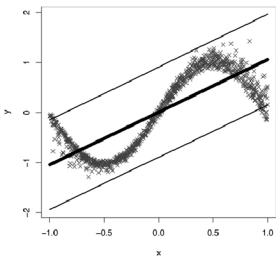

(b) PVNI=0.998

*Slika 6.4*

Prikaz napovednih intervalov z metodama stremljenja in maksimalnega verjetja na problemu sinusoidne funkcije z mešanim šumom in modelom regresijskih dreves, (a) poenostavljena metoda, (b) originalna metoda.




(a) PVNI=0.932




(b) PVNI=0.979

*Slika 6.5*

Napovedni intervali metod stremljenja in maksimalnega verjetja (a) poenostavljena metoda na odsekoma konstantni funkciji in modelu regresijskih dreves, (b) originalna metoda na sinusoidni funkciji z mešanim šumom in modelom linearne regresije.

glede na napovedi modela. Na desni sliki pa vidimo, da se napovedni intervali bolj prilegajo residualom. Čeprav imata pokrivni verjetnosti napovednih intervalov približno enako napako, je razlika relativnih povprečnih intervalov bolj značilna: pri poenostavljeni metodi je RPNI=0.713, pri originalni pa je RPNI=0.302.

Zadnji par slik prikazuje pomanjkljivosti metod stremljenja in maksimalnega verjetja. Na levi sliki 6.5 vidimo dve skupini residualov namesto treh, na desni sliki pa zgolj eno. Lahko vidimo, da metode stremljenja in maksimalnega verjetja stremijo k temu, da zajamejo napovedi in tvorijo pravilne intervale; njihova širina oziroma optimalnost je sekundarnega pomena.

## 6.4 Metodologija testiranja

Pri poskusih z intervalnimi ocenami je metodologija nekoliko drugačna. Stremljenje nudi dobre ocene, skupaj z informacijo stabilnosti rezultatov, vendar stremljenje že izkoriščajo intervalne cenilke zanesljivosti (razdelek 5.1). Zaradi tega vsako podatkovno množico naključno razbijemo na dve enako veliki podmnožici; na učno in testno množico. Učna množica je na razpolago učnim algoritmom regresijskih modelov in metodam ocenjevanja zanesljivosti. Testna množica je namenjena izključno evalvaciji; na njej so izračunane napovedne točnosti, pokrivne verjetnosti napovednih intervalov, relativni povprečni intervali. Zabeleženi so tudi izvajalni časi, da lahko primerjamo učinkovitost posameznih pristopov. Celoten postopek je za stabilnost rezultatov ponovljen 50 krat. V preteklih raziskavah smo preizkusili tudi rabo Studentove distribucije, izkazalo pa se je, da z njo generiramo preširoke napovedne intervale.

Ponovno velja izpostaviti, da pri poskusih ne posvečamo nikakršne pozornosti optimizaciji posameznih modelov, oziroma da je njihova parametrizacija konstantna in nastavljena na povprečne, oz. privzete vrednosti. Pri poskusih obravnavamo modele kot črne škatle, pa naj bodo te dobre ali slabe.

## 6.5 Osnovni rezultati poskusov

*Tabela 6.1*

Eksperimentalna evalvacija intervalnih cenilk zanesljivosti na umetnih podatkovnih množicah. Stolpci PVNI prikazujejo 2.5 percentil, srednjo vrednost in 97.5 percentil distribucije doseženih vrednosti. RPNI in čas izvajanja sta povprečena preko vseh modelov in podatkovnih množic.

| metoda | PVNI2.5 | PVNI50 | PVNI97.5 | RPNI | čas [ms] |
|---|---|---|---|---|---|
| *SMVa* | 0.384 | 0.859 | 1.000 | 0.347 | 4205 |
| *SMVb* | 0.849 | 0.948 | 1.000 | 0.392 | 5070 |
| *NS5* | 0.270 | 0.949 | 1.000 | 0.234 | 396 |
| *NS100* | 0.281 | 0.933 | 1.000 | 0.316 | 1.8 |
| *NSrs* | 0.267 | 0.911 | 0.939 | 0.216 | 102 |
| *KRG* | 0.935 | 0.957 | 0.998 | 0.473 | 348 |

Osnovni rezultati testiranja šestih intervalnih cenilk zanesljivosti na umetnih podatkovnih množicah so prikazani v tabeli 6.1. Prvi trije stolpci predstavljajo distribucijo doseženih pokrivnih verjetnosti napovednih intervalov, natančneje njihov 2.5 percentil, srednjo vrednost in 97.5 percentil. Izbran $\alpha$ je 0.05, zato je tarča vrednosti PVNI 0.95. Glede na to mero je v povprečju najboljša tista metoda, ki ima PVNI50 enak 0.95 ter čim manjši raztros, torej PVNI2.5 in PVNI97.5 čim bližje 0.95. Zadnja dva stolpca predstavljata relativne povprečne napovedne intervale in čas izvajanja posameznih metod.

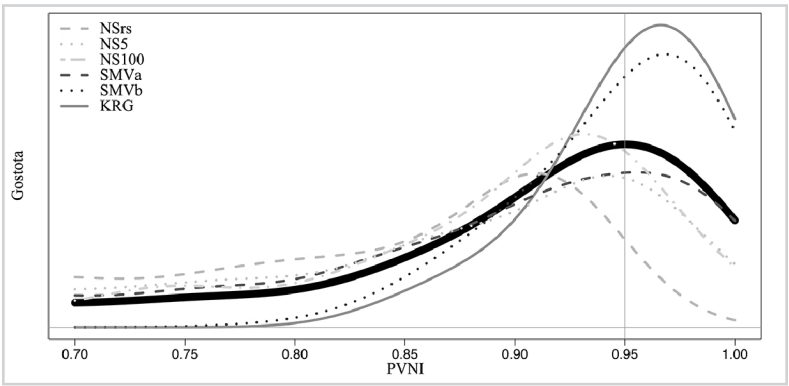
Na realnih podatkovnih množicah, kot je prikazano v tabeli 6.5, metode na osnovi lokalnih okolic *NSrs*, *NS5* and *NS100* podcenjujejo srednjo vrednost PVNI. Na realnih podatkovnih množicah lahko te metode tvorijo preozke intervale, zaradi česar veliko napovedi ni zajetih, to pa se manifestira v nizkih vrednostih PVNI2.5.

Po drugi strani se pa metode stremljenja in maksimalnega verjetja trudijo postaviti pravilne napovedne intervale, ne glede na to, koliko je model slab. To smo videli na sliki 6.5, skoraj identično sliko pa dobimo z metodo *NS100*, za katero v tabeli vidimo

*Tabela 6.2*

Eksperimentalna evalvacija intervalnih cenilk zanesljivosti na realnih podatkovnih množicah. Stolpci PVNI prikazujejo 2.5 percentil, srednjo vrednost in 97.5 percentil distribucije doseženih vrednosti. RPNI in čas izvajanja sta povprečena preko vseh modelov in podatkovnih množic.

| metoda | PVNI2.5 | PVNI50 | PVNI97.5 | RPNI | čas [ms] |
|---|---|---|---|---|---|
| *SMVa* | 0.607 | 0.917 | 1.000 | 0.593 | 466 |
| *SMVb* | 0.852 | 0.967 | 1.000 | 0.822 | 552 |
| *NS5* | 0.323 | 0.851 | 0.960 | 0.435 | 21 |
| *NS100* | 0.354 | 0.875 | 0.990 | 0.503 | 1.4 |
| *NSrs* | 0.340 | 0.785 | 0.918 | 0.383 | 11 |
| *KRG* | 0.852 | 0.961 | 1.000 | 0.733 | 14 |




*Slika 6.6*

Prikaz gostote doseženih vrednosti PVNI. Zastavljena vrednost PVNI 0.95 je označena z navpično črto. Odebeljena črna črta predstavlja porazdelitev vrednosti PVNI vseh metod skupaj. Posamezne metode so označene v legendi.

dobre rezultate. Ko metode ustvarijo pravilne intervale, ki se podatkom ne prilegajo dobro, dobimo bistveno širše napovedne intervale in posledično višje vrednosti RPNI.

Iz obeh tabel je razvidno, da metode na osnovi lokalnih okolic v povprečju postavljajo ožje napovedne intervale (nižje povprečne vrednosti RPNI) in da so časovno mnogo bolj učinkovite od pristopov, ki izkoriščajo stremljenje. Kvantilni regresijski gozdovi se po pravilnosti (doseženih vrednosti PVNI) lahko kosajo z originalno metodo stremljenja in maksimalnega verjetja, tvorijo bolj optimalne napovedne intervale, časovna zahtevnost pa je veliko bolj podobna metodi *NS5*.

Slika 6.6 prikazuje gostote doseženih vrednosti PVNI vseh preizkušenih metod in omogoča njihovo primerjavo. Skupna porazdelitev vrednosti vseh metod je prikazana

*Tabela 6.3*

Odstotki vrednosti PVNI enakih 1.0. Pri teh poskusih napovedni intervali zajemajo vse testne primere.

| *SMVb* | *SMVa* | *KRG* | *NS100* | *NS5* | *NSrs* |
|---|---|---|---|---|---|
| 5.2% | 3.5% | 2.8% | 2.1% | 1.7% | 0% |

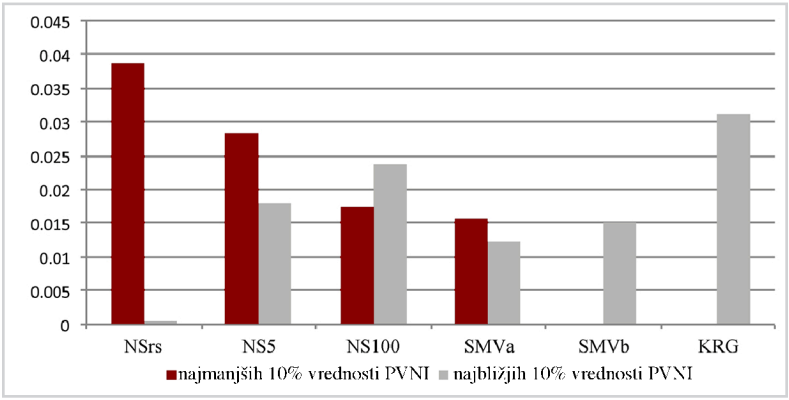
z odebeljeno črto. Vidimo, da v povprečju metode dosežejo različne nizke vrednosti PVNI, vendar pa je vrh gostote vseh vrednosti pri 0.951, kar je izjemno blizu zastavljeni oziroma iskani vrednosti 0.95 (navpična črta).

Metode stremljenja in maksimalnega verjetja so prikazane s temno sivimi črtami, metode z lokalnimi okolicami pa s svetlo sivimi črtami. Iz prikaza je razvidno, da metode *NSrs*, *NS5* and *NS100* v povprečju dosegajo nižje vrednosti PVNI od pričakovane, na drugi strani pričakovane vrednosti pa najdemo metode s stremljenjem in maksimalnim verjetjem ter kvantilni regresijski gozd.

Metoda, ki se je najbližje približala zastavljeni vrednosti PVNI, je *SMVa*, z vrhom pri 0.955, distribucija te metode pa je značilno bolj razpršena od ostalih metod. Drugi najbližji vrh zastavljeni PVNI vrednosti je 0.942, dosegla pa ga je metoda *NS5*. Sledijo vrhovi *KRG* pri 0.967, *SMVb* pri 0.968 in analitičnega pristopa *NS100* pri 0.931. Vrh, najbolj oddaljen od zastavljene vrednosti, pripada metodi *NSrs*, pri 0.910.

Kadar napovedni intervali zajemajo celotni prostor odvisne spremenljivke, metoda doseže pokrivno verjetnost napovednih intervalov 1.0. Odstotki teh vrednosti PVNI enakih 1.0 so prikazani v tabeli 6.3, v kateri med drugim vidimo, da metoda z razvrščanjem v skupine ni nikdar storila te napake. Na drugi strani pa sta največ grešili metodi stremljenja in maksimalnega verjetja: metoda *SMVb* je zajela vse testne primere v 5.2% poskusov, metoda *SMVa* pa v 3.5%. Ob ignoriranju teh primerov v distribuciji vrednosti PVNI, najbližji PVNI zadanemu še zmeraj pripada metodi *SMVa* pri 0.947, vendar je zdaj druga najbližja vrednost 0.941 metode *NS5*. Ne tako daleč sledijo metoda *KRG* z vrhom pri 0.965, metoda *NS100* pri 0.933 in *SMVb* pri 0.966. PVNI metode *NSrs* se ne spremeni (0.910), vendar ostaja najdlje od zastavljene vrednosti PVNI.

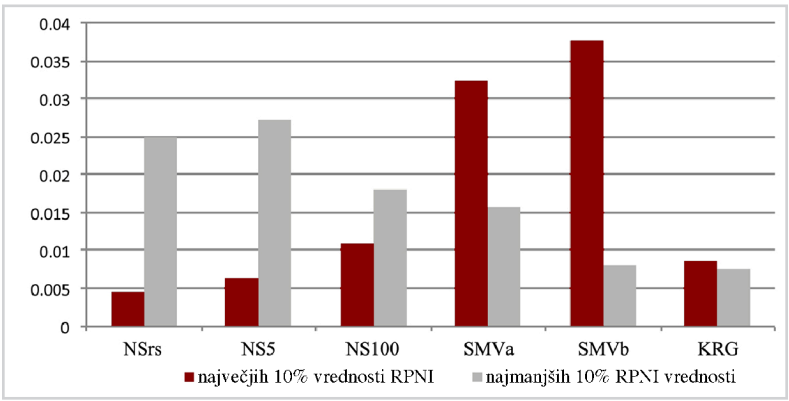
Slika 6.7 prikazuje, kako je med posameznimi pristopi porazdeljenih 10% najnižjih vrednosti PVNI in 10% vrednosti PVNI najbližjih zastavljeni vrednosti. Iz tega vidika deluje metoda *NSrs* najbolj riskantno in *KRG* najbolj varno. Zanimivo je, da analitični napovedni intervali *NS100* tu dajejo vtis, da so boljši od ostalih metod. Vendar smo že omenili, da so tudi ti intervali v povprečju pravilni, ampak se le redko dobro prilegajo




*Slika 6.7*

Porazdelitev poskusov (njihov odstotek) z najboljšimi in najslabšimi dosežennimi vrednostmi PVNI posameznih pristopov.

podatkom.




*Slika 6.8*

Porazdelitev poskusov (njihov odstotek) z najboljšimi in najslabšimi dosežennimi vrednostmi RPNI posameznih pristopov.

Zdaj je očitno, da je pomembno upoštevati in primerjati tudi relativne povprečne napovedne intervale. Če dve metodi dosežeta enako pokrivno verjetnost napovednih intervalov, potem ima tista, ki ima nižji RPNI, ožje intervale, zaradi česar bi veljala za boljšo. Slika 6.8 prikazuje, katere metode so dosegle najnižje (najboljše) in najvišje (najslabše) vrednosti RPNI. Nizke vrednosti RPNI so večinoma dosegle metode lokalnih okolic zaradi njihove narave tvorjenja optimalnejših intervalov. Na podatkovni množici *servo* so metode *NSrs*, *NS5*, in *NS100*, v navezi z modelom umetne nevronske

mreže, dosegle pokrivno verjetnost 0, relativni povprečni intervali pa se gibljejo med 0.0015 in 0.002. Največje vrednosti RPNI so dosegle metode stremljenja in maksimalnega verjetja. Na podatkovni množici *triazines* in modelom umetne nevronske mreže je metoda *SMVa* dosegla vrednost RPNI 5.3, pripadajoča vrednost PVNI pa je enaka 1.0. To pomeni, da je metoda zajela vse testne primere na račun zelo širokih napovednih intervalov. Sodeč po sliki je pristop *KRG* dosegel najmanj ekstremno nizkih vrednosti RPNI in tudi relativno malo največjih vrednosti in je tudi s tega stališča, v povprečju, najbolj varna metoda za uporabo.

# *Naprednejši pristopi z intervalnimi cenilkami*

7

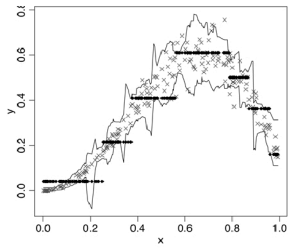
V tem poglavju se posvečamo naprednejšim pristopom, ki temeljijo na intervalnih cenilkah, predstavljenih v 5. poglavju. Osredotočamo se na pristope, ki so se v prejšnjem poglavju izkazali za uspešnejše. Zaradi boljše preglednosti se slabšim pristopom, metodam *SMVa*, *NS100* in *NSrs* ne posvečamo več. Za primerjavo je v tem poglavju izbrana $\alpha$=0.1, torej želimo 90% napovedne intervale. S tem dovolimo, da je manj pravih vrednosti vsebovanih znotraj intervalov, kar povzroči da so napake v prileganju podatkom bolj izrazite.

Poglavje se začne z vpeljavo kombinirane metode, ki združuje optimalne in pravilne intervale. Sledijo vizualizacijske tehnike, ki omogočajo primerjavo intervalov in primerjavo modelov. Poglavje zaključujeta študiji izbire napovedi s pomočjo združevanja modelov in izbire modela s statistiko, ki združuje obe meri uspešnosti napovednih intervalov.

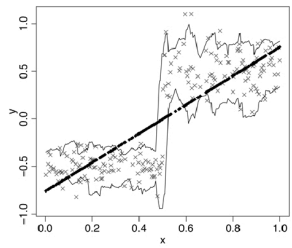
## 7.1 Združevanje optimalnih in pravilnih intervalov

V prejšnjem poglavju smo videli, da metode na osnovi lokalnih okolic tvorijo bolj optimalne napovedne intervale in da metode na osnovi stremljenja in maksimalnega verjetja tvorijo bolj pravilne napovedne intervale.

Za poljubni model pričakujemo, da je razlika med pravilnimi in optimalnimi napovednimi intervali čim manjša. Tukaj postavljamo hipotezo, da za najboljši model velja, da imajo napovedni intervali po pristopu stremljenja in maksimalnega verjetja pravilno globalno sliko in da so napovedni intervali lahko celo bolj optimalni od napovednih intervalov metod na osnovi lokalnih okolic.



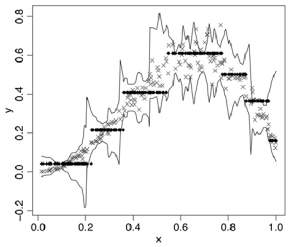

(a) PVNI=0.90, RPNI=0.26



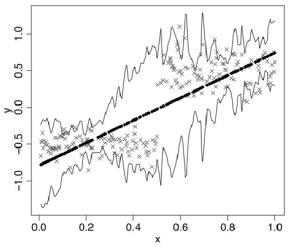

(b) PVNI=0.85, RPNI=0.36

*Slika 7.1*

Primera optimalnih napovednih intervalov z metodo *NS5* za (a) regresijsko drevo na nelinearni funkciji s heterogenim šumom in (b) linearno regresijo na odsekovno konstantni funkciji z različnima stopnjama šuma.



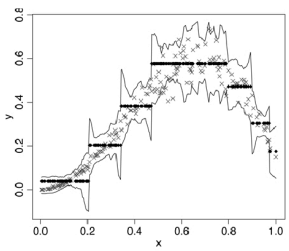

(a) PVNI=0.89, RPNI=0.30



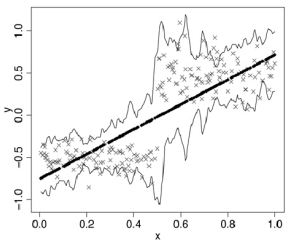

(b) PVNI=0.91, RPNI=0.63

*Slika 7.2*

Primera pravilnih napovednih intervalov z metodo *SMVb* na istih problemih kot na sliki 7.1




(a) PVNI=0.94, RPNI=0.28




(b) PVNI=0.91, RPNI=0.47

*Slika 7.3*

Napovedni intervali kombinirane metode *NS-SMV* na istih problemih kot na slikah 7.1 in 7.2.

Kombinirani napovedni intervali *NS-SMV* se enostavno izračuna. Spodnja in zgornja meja napovednega intervala sta povprečje spodnjih, oziroma zgornjih mej intervalov metod *NS5* in *SMVb*. Če z $\top$ označimo zgornjo mejo in z $\bot$ spodnjo mejo, potem velja

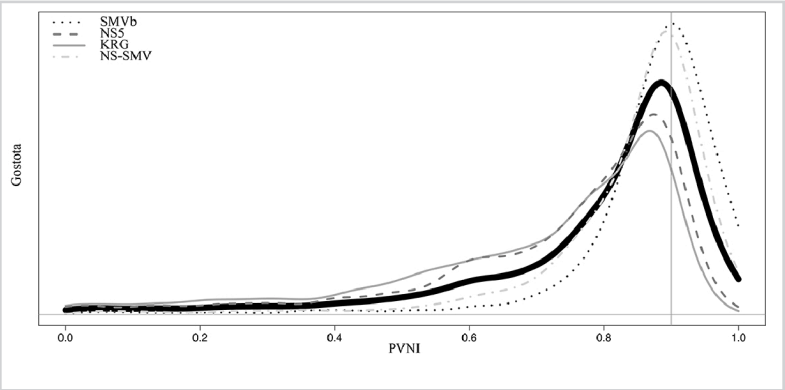
$$NS\text{-}SMV\top = \frac{NS5\top + SMVb\top}{2} \text{ in } NS\text{-}SMV\bot = \frac{NS5\bot + SMVb\bot}{2}.$$

S povprečenjem se tudi odpovedujemo garanciji vsebovanosti napovedi modelov, katero nosijo metode na osnovi stremljenja, vendar pridobimo prilagodljivost lokalnim okolicam. Za demonstracijo uporabimo regresijska drevesa na problemu nelinearne funkcije s heterogenim šumom in linearno regresijo na odsekovno konstantni funkciji z različnima stopnjama šuma. Slika 7.1 prikazuje optimalne napovedne intervale (z metodo *NS5*), slika 7.2 pa pravilne (z metodo *SMVb*). Slika 7.3 prikazuje kombinirane napovedne intervale, za katere vidimo, da se intervali bolje prilegajo podatkom, da so do neke mere bolj stabilni od intervalov posamičnih metod in da so tudi mere uspešnosti boljše.

Sliki 7.4 in 7.5 prikazujeta porazdelitvi gostote doseženih pokrivnih verjetnosti na umetnih in realnih podatkovnih množicah posebej. Iz ožjega nabora osnovnih metod so napovedni intervali *NS5* prikazani s sivo črtkano črto, stremljenja in maksimalnega verjetja s svetlo sivo pikčasto črto in kombinirane metode s pikčasto črtkano črto. Pri umetnih podatkovnih množicah je očitno, da so napovedni intervali vseh pristopov mnogo bližje zastavljeni verjetnosti 90% (označeno z navpično črto) kot pri realnih množicah. Iz premika distribucij vrednosti PVNI je očitno, da so realne podatkovne množice trši oreh, a kljub temu sta razvrstitvi vrhov distribucij posameznih metod enaki. Po pričakovanjih so najbližji intervali metode stremljenja in maksimalnega verjetja. Kombinirani napovedni intervali *NS-SMV* so drugi najbližji, nato sledi metoda *NS5* in najdlje so intervali metode *KRG*.

## 7.2 Vizualizacijske tehnike

Prikazi napovednih intervalov, ki smo jih uporabljali do sedaj, odlično prikazujejo delovanje posameznih metod, vendar je njihova uporabnost zelo omejena. Pomanjkljivost je ta, da jih lahko uporabljamo le za prikaz dvorazsežnih podatkov, odvisne spremenljivke in enega atributa. Sicer bi jih lahko uporabljali tudi za več razsežnosti, a hitro postanejo nepraktični. V nadaljevanju razdelka povzemamo tehniko za vizualno pri-




*Slika 7.4*

Prikaz gostote dosežen-ih vrednosti PVNI na umetnih podatkovnih množicah. Zastavljena vrednost PVNI 0.90 je označena z navpično črto. Odebeljena črna črta predstavlja porazdelitev vrednosti PVNI vseh metod skupaj. Posamezne metode so označene v legendi.

merjavo intervalov različnih metod in predstavljamo tehniko za primerjavo ustreznosti modelov.

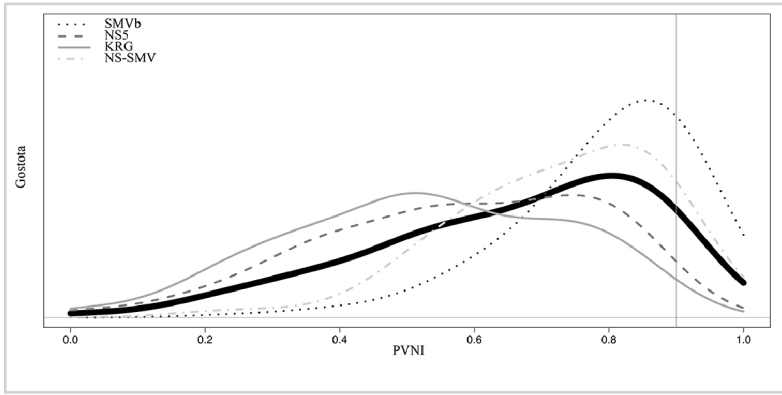
### *7.2.1 Primerjava intervalov*

V [44] lahko najdemo preprosto in intuitivno tehniko vizualiziranja napovednih intervalov, ki omogoča primerjavo in evalvacijo napovednih intervalov več metod na neki problemski množici.

Testni primeri so najprej urejeni po naraščajoči dolžini napovednih intervalov. Za sredinsko poravnanost se povprečja zgornjih in spodnjih mej napovednih intervalov odštejejo od vseh pravih vrednosti odvisne spremenljivke in napovednih intervalov. Na abscisni osi so torej testni primeri, na ordinatni pa poravnani napovedni intervali in prave vrednosti odvisne spremenljivke.

Slika 7.6 ilustrira to vizualizacijsko tehniko na sinusoidni funkciji z mešanim šumom; trije prikazi na levi strani pripadajo linearni regresiji in trije na desni strani umetni nevronski mreži. Pri linearni regresiji je za metodo stremljenja in maksimalnega verjetja ponovno očitno, da si metoda prizadeva ustvariti pravilne napovedne intervale ne glede na njihovo dolžino.

Vizualizacijska tehnika skriva napovedi modelov, tako da popolnoma napačne napovedi linearne regresije niso opazne, kakor so bile na prejšnjih prikazih. Vseeno je precej očitno, da so napovedni intervali v desnem stolpcu krajši in da se bolje prilegajo podatkom. Kljub pomanjkljivostim vizualizacijske tehnike bi na osnovi prikazov raje




*Slika 7.5*

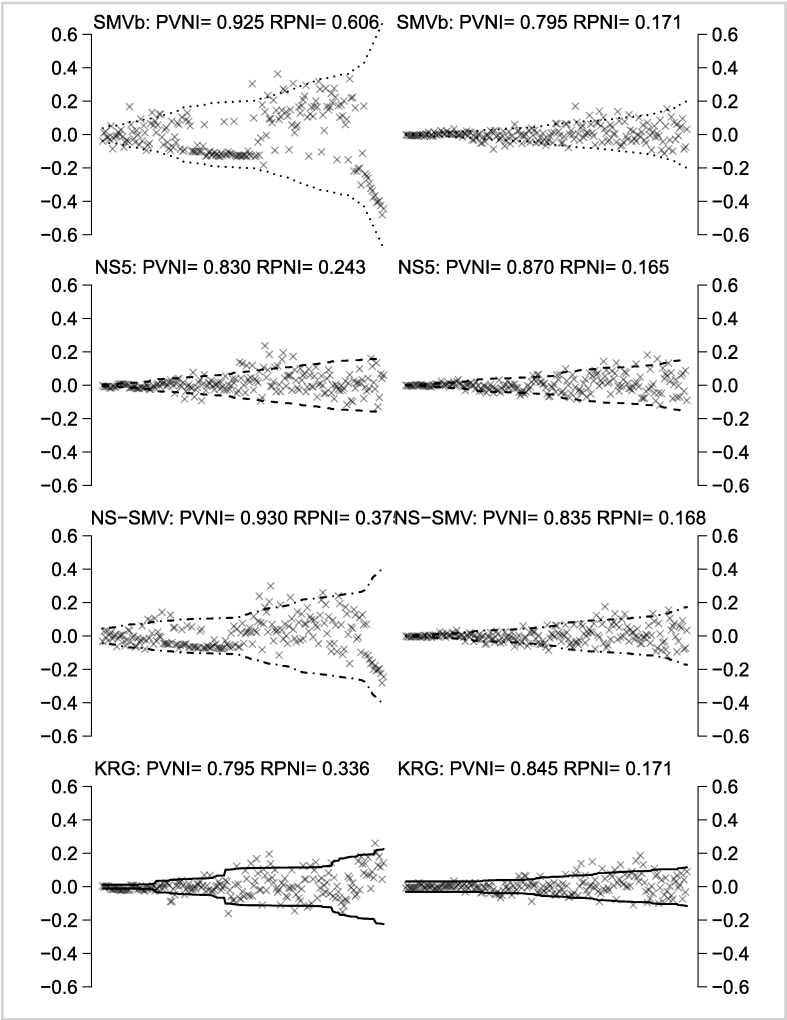
Prikaz gostote doseženih vrednosti PVNI na realnih podatkovnih množicah. Zastavljena vrednost PVNI 0.90 je označena z navpično črto.

izbrali umetno nevronsko mrežo kot pa linearno regresijo. Ob upoštevanju prileganja vrednostim odvisne spremenljivke in vrednosti PVNI ter RPNI, so napovedni intervali metode *NS5* najbolj točni in informativni.

### *7.2.2 Primerjava modelov*

Druga, zelo koristna vizualizacijska tehnika je prikaz residualov proti napovedanim vrednostim [55]. Pri teh prikazih napovedane vrednosti tvorijo abscisno os, ordinatna os pa predstavlja velikost napake oz. residuale. Residuali bi morali biti naključno porazdeljeni in boljši kot je model bližje so residuali abscisni osi. Po navadi je zelo preprosto prepoznati neprimerne modele, katerih napovedi neustrezno pokrivajo prostor odvisne spremenljivke.

Za ilustracijo ponovno uporabljamo problem sinusoidne funkcije z mešanim šumom. Slika 7.7 prikazuje residuale v odvisnosti od napovedanih vrednosti za vse regresijske modele, podaja pa še koren srednje kvadratne napake (RMSE) in relativni koren srednje kvadratne napake (rRMSE). Vodoravne črte predstavljajo distribucijo residualov in prikazujejo, kako so residuali porazdeljeni okoli ničle. Napovedni intervali na teh vizualizacijah prispevajo to, da njihov raztros, medsebojna nesoglasja (križanja) in posamezne deviacije govorijo v prid manj zanesljivim napovedim, oziroma identificirajo modele, nagnjene k napačnim napovedim. Za naš primer se zelo dobro vidi, da modeli bagging, regresijska drevesa, metoda podpornih vektorjev in linearna regresija ne ustrezajo, oziroma se njihove napovedi ne prilegajo podatkovni množici. Umetne

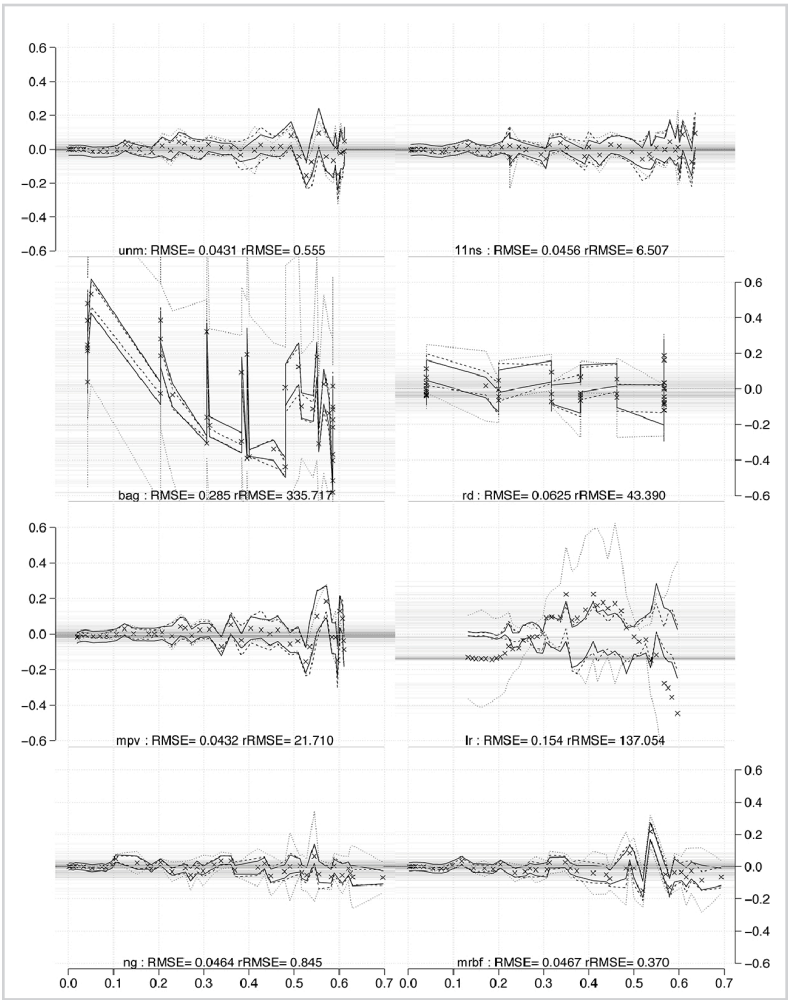



*Slika 7.6*

Urejeni in poravnani napovedni intervali in pripadajoče vrednosti odvisne spremenljivke na problemu sinusoidne funkcije z mešanim šumom. Levi stolpec prikazuje napovedne intervale za model linearne regresije in desni za umetne nevronske mreže.

nevronske mreže, najbližji sosedi, naključni gozd in mreža radialnih baznih funkcij se dobro prilegajo podatkom in pravzaprav ni očitno, kateri model je najboljši.

Na teh prikazih so tudi osamelci hitro razvidni. Morda bi najboljše rezultate dobili s kombinacijo omenjenih modelov, na primer če uporabimo napovedi mreže radialnih baznih funkcij na enem delu domene (levi del grafa) in napovedi naključnega gozda na drugem delu (desni del grafa). Vendar to je le ugibanje, na podlagi vizualizacije. Izbiro posameznih napovedi na osnovi napovednih intervalov obravnavamo v naslednjih razdelkih.




*Slika 7.7*

Prikaz residualov proti napovedanim vrednostim za sinusoidno funkcijo z mešanim šumom. Residuali so označeni s križci; črte, ki predstavljajo napovedne intervale, so za *SMVb* pikčaste, za *NS5* črtkane in za *KRG* polne. Vodoravne črte predstavljajo distribucijo residualov in prikazujejo, kako so residuali porazdeljeni okoli ničle.

## 7.3 Izbira napovedi z agregacijo modelov

Med najbolj zgodnjimi načini združevanja različnih modelov strojnega učenja je bil sistem, ki po principu deli-in-vladaj razdeli atributni prostor s pomočjo dveh ali več klasifikatorjev [57]. Metoda bagging je tudi ena izmed zgodnjih, intuitivnih in morda najbolj preprostih načinov združevanja modelov, vendar deluje presenetljivo dobro.

Izbor modelov je morda glavni razlog, zakaj se v praksi uporabljajo ti pristopi. Kombiniranje modelov ne zagotavlja izboljšanja točnosti napram najboljšemu izmed modelov ansambla, vendar zagotovo zmanjša celokupno tveganje slabih napovedi.

Naš pristop uporablja širino napovednih intervalov za izbor napovedi. Naš osnovni pristop uporablja za združevanje funkcijo minimum. To pomeni, da izbere napoved modela, katere napovedni interval je najkrajši.

Druga inačica uporablja uteževanje. Najprej poiščemo najkrajši in najdaljši napovedni interval za posamezni testni primer, označimo ju z $\top$ in $\bot$. Dolžino najkrajšega napovednega intervala odštejemo od vseh napovednih intervalov (označimo jih z $\{I_1, \ldots, I_m\}$), ki so za tem normalizirani z najdaljšim napovednim intervalom. Uteži dobimo tako, da te vrednosti odštejemo od 1, zato da najširši interval nima doprinosa, najkrajši interval pa ima utež 1. Množico vseh uteži ponovno normaliziramo, tako da je njihova vsota enaka 1. Utež $u_i$ lahko zapišemo kot

$$u_i = \frac{1 - (I_i - \bot)/(\top - \bot)}{\sum_i 1 - (I_i - \bot)/(\top - \bot)}.$$

Končno napoved tvori utežena vsota napovedi posameznih modelov, napovedni interval pa je oblikovan na enak način (utežena vsota zgornjih mej in utežena vsota spodnjih mej).

## 7.4 Kombinirana statistika

Po navadi je težko istočasno primerjati več statistik preko več metod, domen in modelov. Da se izognemo tej komplikaciji, tukaj predlagamo kombinirano statistiko, ki zajema obe meri uspešnosti napovednih intervalov. V [58] smo preizkusili uporabo zgolj dolžine napovednih intervalov za izbiro napovedi. Na umetnih podatkovnih množicah so bili poskusi obetavni, vendar so realne množice hitro zavrnile to preprosto hipotezo.

Že vemo, da če imata dve metodi enake pokrivne verjetnosti napovednih intervalov, je tista z manjšim relativnim povprečnim napovednim intervalom boljša izbira. Intuitivno je jasno, da ima najboljša metoda vrednost PVNI čim bližje zastavljeni verjetnosti in da ima čim manjšo vrednost RPNI. Tu trdimo, da je optimalnost težje dosegljiva od pravilnosti, saj je trivialno doseči vrednost PVNI 1, medtem ko je težko doseči nizko vrednost RPNI. Zaradi tega pri naši kombinirani oceni dajemo večji vpliv meri RPNI. Osnovna enačba kombiniranje statistike je

$$100 \cdot \mathrm{RPNI} + \log\left((\mathrm{PVNI}^* - \mathrm{PVNI})^2\right),$$

kjer je $\mathrm{PVNI}^*$ zastavljena pokrivna verjetnost napovednih intervalov. Ampak ker se logaritem približuje minus neskončno, ko se vrednost PVNI približuje zastavljeni vrednosti, dodatno omejimo ta doprinos. Končna oblika kombinirane statistike RPNI-PVNI je tako

$$100 \cdot \mathrm{RPNI} + \log\left(\left(\max(\mathrm{PVNI}^* - \mathrm{PVNI})^2\right), 10^{-10}\right). \tag{7.1}$$

## 7.5 Rezultati

Slika 7.8 predstavlja glavne rezultate tega poglavja in prikazuje toplotni zemljevid kombinirane statistike RPNI-PVNI (predstavljene v prejšnjem razdelku) za vse umetne in realne podatkovne množice (po vrsticah) za vse kombinacije modelov in napovednih intervalov (po stolpcih). Vključuje tudi oba pristopa združevanja napovedi (predstavljena v razdelku 7.3), primeri predznačeni z *u* so uteženi.

Nekateri poskusi so dosegli ekstremno visoke vrednosti (zaradi visoke statistike RPNI), nekatere pa negativne (zaradi statistike PVNI zelo blizu zastavljene vrednosti), zato vizualizacija uporablja obratne vrednosti logaritmov, s čimer postane prikaz razločen. Boljše, torej manjše vrednosti kombinirane statistike so predstavljene z modro, slabše, oziroma večje pa z rdečo barvo.

Vrstice in stolpci so urejeni po naraščajočih kumulativnih vsotah, kar omogoča primerjavo težavnosti podatkovnih množic in uspešnosti posameznih pristopov. Umetne podatkovne množice imajo za imenom funkcije označeno število naključnih atributov (N3 pomeni da so trije dodatni, šumni atributi), za tem pa oznako velikosti podatkovne množice od 1 do 6 oziroma od 50 do 1600, kjer je vsaka naslednja množica dvakrat večja.

Od zgoraj navzdol lahko razberemo, kako so v povprečju težavne posamezne podatkovne množice. Na sliki je očitno, da v zgornjem delu prevladujejo umetne podatkovne množice in da realne podatkovne množice prevladujejo na spodnjem delu. Kljub temu so množice dobro razpršene, saj se prva realna množica nahaja že na četrtem mestu.

Ob preučevanju doseženih vrednosti na umetnih podatkovnih množicah lahko vidimo, katere kombinacije modelov in metod izstopajo. En tak primer najdemo pri linearni funkciji s heterogenim šumom, kjer lahko razberemo, da se model bagging v kombinaciji z metodo *SMVb* ali *NS-SMV* slabo prilega podatkom. Če bi gledali prikaz zgolj vrednosti PVNI, tega ne bi opazili. Ta dva primera sta tudi najslabša izmed vseh prikazanih. Med boljšimi statistikami je nekoliko težje določiti absolutno najboljše, saj so razlike po navadi precej majhne. V prvi vrstici pa za podatkovno množico 800 primerov linearne funkcije z linearno odvisnim šumom in enim dodatnim šumnim atributom lahko razberemo, da se najbolje prilegajo napovedi linearne regresije in napovedni intervali kvantilnega regresijskega gozda, druga najboljša kombinacija pa so napovedi 11. najbližjih sosedov z napovednimi intervali metode stremljenja in maksimalnega verjetja. Med umetnimi podatkovnimi množicami so se izkazale za najzahtevnejše tiste, ki so vzorčene iz nezveznih funkcij (odsekoma konstantna funkcija – *pieceConst* in odsekoma linearna – *linearPiece*) ter sinusiodne funkcije s heterogenim šumom (*nonlinearHetero*).

V razvrstitvi seštevkov kumulativnih statistik po kombinacijah modelov in napovednih intervalov ter pristopih združevanja napovedi, torej pri branju slike od leve proti desni, se zelo jasno vidi, da so metode z lokalnimi okolicami na levi ter metode stremljenja in maksimalnega verjetna na desni. Ker je manjše boljše, ta slika ponovno kaže na to, da se metode na osnovi lokalnih okolic v povprečju praviloma bolje prilegajo podatkom. Zelo pozitivno je to, da se pristopi združevanja napovedi nahajajo večinoma na levi strani, kjer je združevanje z metodo *NS5* v povprečju najboljše in tudi kombinirana metoda *NS-SMV* je boljša od pristopov s stremljenjem in maksimalnim verjetjem. Lahko tudi razberemo, da postopki združevanja napovedi v povprečju dosegajo boljše vrednosti kombinirane statistike, kakor jih posamezni modeli. Po pričakovanjih se le redko zgodi, da združene napovedi dajo boljše rezultate kakor najboljši posamezni model, vendar so v povprečju dosti boljše od slabih posameznih modelov.

Idealno bi bilo, če bi nam kombinirana statistika nudila informacije o pričakovani napaki, ki bo dosežena. Tabela 7.1 prikazuje Spearmanove korelacijske koeficiente med doseženimi srednjimi kvadratičnimi napakami ter kombiniranimi statistikami. Z izje-

mo metode najbližjih sosedov (*NS5*) se dosežene stopnje korelacije gibljejo med 0.5 in 0.7. To pomeni, da je kumulativna statistika informativna do določene stopnje, vendar se ne moremo zanesti, da bo vedno izbrala kombinacijo z najmanjšo napako. Združevanje pristopov *NS5* ali *KRG* z uporabo minimuma generira napovedne intervale, ki so nepristranski.

Na koncu si še oglejmo rezultate pristopov združevanja posameznih napovedi. Vprašanje je, ali so pristopi zmožni generirati manjšo kumulativno napako od posameznih modelov. Izmed 1248 možnih izidov so pristopi združevanja dosegli značilno manjšo kvadratno napako v 181 primerih, kar predstavlja 15% testov. Tabela 7.2 prikazuje, kako so ti pozitivni primeri razporejeni med pristopi. Očitno je, da uteženo združevanje deluje mnogo bolje kakor uporaba minimuma. Zanimivo je, da je slednje izboljšalo napovedi le v desetih primerih, uteženo združevanje pa v preostalih 171 primerih.

*Tabela 7.1*

Spearmanovi korelacijski koeficienti med RMSE in statistiko PVNI-RPNI.

| unm | 11ns | bag | rd |
|---|---|---|---|
| 0.705 | 0.689 | 0.629 | 0.708 |
| mpv | lr | ng | mrbf |
| 0.660 | 0.691 | 0.539 | 0.739 |

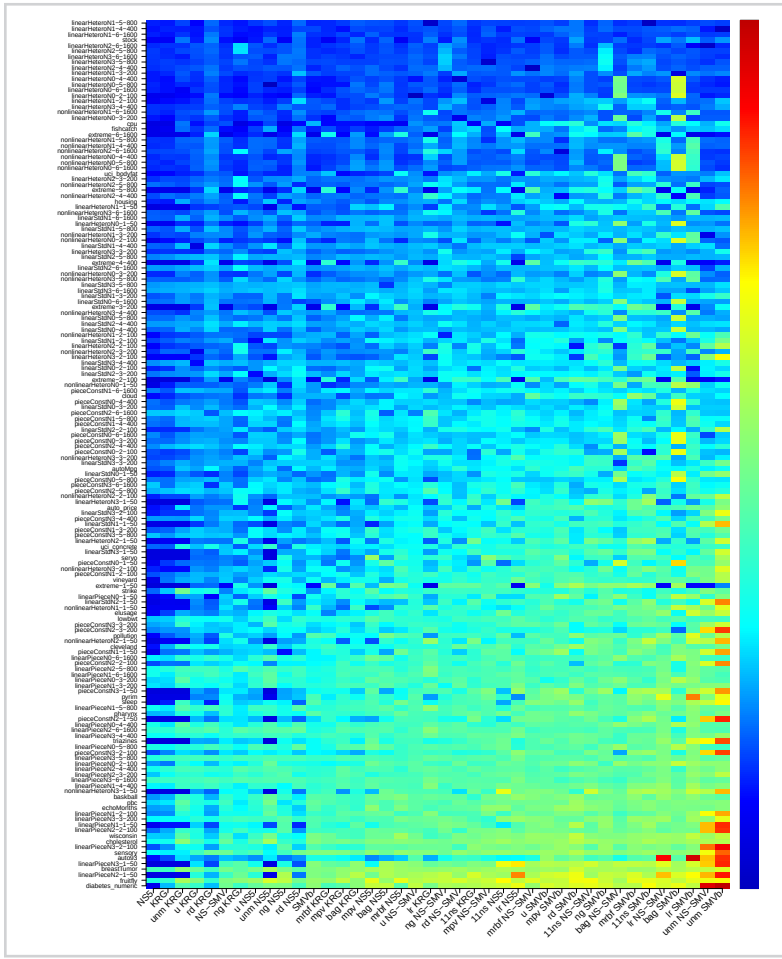
| *NS5* | *SMVb* | *KRG* | *NS-SMV* |
|---|---|---|---|
| -0.0868 | 0.597 | 0.0552 | 0.538 |

| u*NS5* | u*SMVb* | u*KRG* | u*LN-SMV* |
|---|---|---|---|
| 0.491 | 0.627 | 0.472 | 0.609 |

*Tabela 7.2*

Odstotek primerov, v katerih so pristopi združevanja dosegli manjši RMSE kot katerikoli posamezni model.

| *NS5* | *SMVb* | *KRG* | *NS-SMV* | u*NS5* | u*SMVb* | u*KRG* | u*NS-SMV* |
|---|---|---|---|---|---|---|---|
| 0 | 3.8% | 0 | 2.5% | 22% | 30% | 25% | 31% |



*Slika 7.8*

Prikaz vseh kombiniranih statistik RPNI-PVNI. Rdeča predstavlja največje vrednosti statistike in modra predstavlja najmanjše — manjše je boljše. Vrstice in stolpci so urejeni po naraščajočih kumulativnih vsotah. Slika je dostopna na http://lkm.fri.uni-lj.si/dp78.pdf.

# *Zaključki*

*8*

Prvotni cilj je bil odgovoriti na vprašanje, ali je možno razviti splošno metodo ocenjevanja zanesljivosti posameznih napovedi pri nadzorovanem učenju in seveda kako. Tu je mišljena metoda, ki v povprečju nudi dobre rezultate na poljubnih učnih množicah in/ali poljubnih modelih. Kratek odgovor je, da ne obstaja ena splošna metoda, daljši odgovor pa predstavlja doktorsko delo, v katerem obravnavamo točkovne in intervalne cenilke zanesljivosti posameznih napovedi.

Predstavili smo referenčno oceno zanesljivosti, vendar je ta omejena na klasifikacijske probleme. Referenčna ocena temelji na dejstvu, da imamo pri klasifikaciji opravka z binarnim problemom (napoved je pravilna, napoved ni pravilna), katerega lahko verjetnostno kvantificiramo in izpeljemo cenilko pričakovane napake. S pomočjo referenčne ocene smo opravili novo primerjalno študijo uporabnosti točkovnih cenilk na klasifikacijskih primerih. Analiza z uporabo referenčne ocene je pokazala, da so točkovne ocene zanesljivosti koristne le v redkih primerih. Gotovo so koristne na primer takrat, ko smo omejeni na uporabo ne optimalnih modelov (na primer zaradi časovne zahtevnosti ali kakega drugega razloga), preprosti modeli točkovnih cenilk pa dobro ujamejo dani problem. Odkrili smo tudi določene pomanjkljivosti nekaterih metod. Naš naivni pristop na osnovi gostote se je izkazal za popolnoma neuspešnega, bolj smiselno bi bilo zasnovati cenilko, ki bi zanesljivost ocenjevala glede na odvisno spremenljivko — pri klasifikaciji poznamo cenilko ki meri razdaljo od klasifikacijskega roba oziroma glede na ocene težavnosti klasifikacije posameznih primerov [59].

V preteklosti še ni bilo opravljene primerjave intervalnih cenilk. V tem delu smo jih postavili na skupni imenovalec ter tako omogočili njihovo primerjavo. Analiza je razkrila dualno naravo intervalnih cenilk, v smislu da družina metod na osnovi stremljenja in maksimalnega verjetja optimizira pravilnost, druga družina metod na osnovi lokalnih okolic pa optimalnost. Zato smo vpeljali kombinirano metodo, ki združuje lastnosti obeh družin. Izkazalo se je, da kombinirani pristop nudi bolj robustne napovedne intervale. Metoda ki temelji na razvrščanju v skupine se v naših poskusih ni obnesla pretirano dobro, saj so napovedni itervali preozki in se preveč prilegajo podatkom. Zato se lahko v nadaljnjem delu obravnava drugačne pristope, na primer hierarhično razvrščanje oz. razvrščanje s principom vzdržnosti [60].

Primerni grafični prikaz je pomemben in zelo koristen za razvoj uporabnikove intuicije in njegovega razumevanja. Obstoječe statistike, ki nudijo informacije o povprečni uspešnosti modelov, niso dovolj informativne, zato smo, poleg vizualizacijske tehnike za primerjavo napovednih intervalov več pristopov, predstavili novo vizualizacijsko

tehniko. Ta omogoča odkrivanje zakonitosti v podatkih ter vizualno primerjavo in evalvacijo več modelov.

Na osnovi statistik uspešnosti intervalnih ocen smo predlagali postopek oziroma novo kombinirano statistiko, s katero je moč robustno izbirati in združevati regresijske napovedi. Novi način ocenjevanja modelov in združevanja posameznih napovedi modelov nudi prilegajoče in posledično bolj zanesljive napovedi.

## 8.1 Razprava in nadaljnje delo

Raziskave smo začeli z zelo splošnim ciljem, za katerega se je sčasoma izkazalo, da v celoti ni dosegljiv. Kot stranski produkt iskanja odgovora na izhodiščno vprašanje smo dobili zaokrožen pregled področja ocenjevanja zanesljivosti pri nadzorovanem učenju. Nadzorovano učenje se zaradi inherentnih razlik med diskretnim in zveznim deli na klasifikacijo in regresijo. Taiste razlike preprečujejo prenos klasifikacijske referenčne ocene na regresijo in tudi preprečujejo uporabo intervalnih cenilk pri klasifikaciji.

Intervalne cenilke zanesljivosti posameznih napovedi so se izkazale za intuitivne, vizualno atraktivne in zlahka doumljive. Pojasnili smo, da ni naravne preslikave v diskretni svet klasifikacije. Vendar obstajajo tudi drugačni pristopi, na primer ogrodje konformnih napovedi [61] v strojnem učenju, ki z določeno gotovostjo vrača množico možnih razredov in zametki simboličnih podatkov [62] v statističnih krogih, kjer so vsi podatki podani z intervali.

Časovne zahtevnosti algoritmov strojnega učenja lahko z večanjem števila podatkov hitro postanejo neobvladljive. Testiranje in ponavljanje meritev časovno kompleksnost le še poveča. Naš izbor testnih množic morda deluje arhaično, ostredotočili pa smo se tudi na klasične, dandanes morda nekoliko zastarele, algoritme strojnega učenja. Kar se števila atributov tiče, nas implementacije nekaterih modelov navzgor omejujejo na 32 kategoričnih atributov (npr. naključni gozd). Kar se pa tiče števila učnih primerov, časovna kompleksnost raste hitreje kot prostorska in zastavljena metodologija preseže vse razumne meje že pri nekaj tisoč primerih. Poskusi so izvajani na računalniku z 2.8GHz procesorjem Intel Core i7, in v prikazane rezultate je bilo vloženih več procesorskih tednov. Pri iskanju zgornje meje uporabnega števila primerov hitro odkrijemo, da so tudi nekatere implementacije algoritmov strojnega učenja že same po sebi časovno zahtevne. Ob poskusu uporabe podatkovne množice *sat*, ki ima 36 atributov in 4435 učnih ter 2000 testnih primerov, smo odkrili kar nekaj omejitev. Na primer gradnja optimalnega odločitvenega drevesa z 2000 primeri (testna množica) lahko merimo v

sekundah, pri 4435 primerih (učna množica) pa lahko zahtevnost merimo kar v dneh.

Teoretične ocene časovne zahtevnosti imajo za praktično posledico to, da z večjimi količinami podatkov posamezne metode postanejo praktično neuporabne. Najzahtevnejša točkovna cenilka je lokalno prečno preverjanje in potreben čas je močno odvisen od časovne zahtevnosti algoritma strojnega učenja; za občutek: pri prej omenjeni podatkovni množici *sat*, traja izračun ocen LCV z uporabo hitre implementacije najbližjih sosedov dobre 3 ure, pri uporabi z nevronsko mrežo pa linearna ekstrapolacija pravi, da bi na rezultate morali čakati kar 120 ur. Naslednja najzahtevnejša metoda temelji na stremljenju in uporabnost takšnih pristopov je ravno tako močno odvisna od časovne zahtevnosti izbranega algoritma strojnega učenja.

Predstavljene intervalne cenilke so v praksi računsko manj zahtevne, časovno je najbolj zahteven postopek stremljenja. Za primerjavo, analiza regresijske podatkovne množice *parkinson UPDRS* s 16 atributi in 5875 primeri po naši metodologiji (vsi modeli, eno izvajanje) porabi slabi 2 uri. Pri večjih množicah, npr. s podatkovno množico *slice localization*, ki ima 383 numeričnih atributov in 53500 primerov, pa že zaidemo v prostorske težave (prvi model, ki preseže kapacitete testnega računalnika je model bagging).

Testiranja so pokazala, da poleg kombinirane metode tudi kvantilni regresijski gozd v povprečju daje dobre, stabilne in kredibilne napovedne intervale. Vendar je to le en model izmed novejše, pestre družine modelov, katerih cilj je učenje kvantilov, oziroma še nekoliko širše družine učenja pogojnih porazdelitev gostote.

Izkoriščanje oz. vpeljava naključja se pri različnih pristopih izkaže za ugodno, primeri tega so stremljenje, naključni gozdovi, kvantilni regresijski gozdovi, itd. Zato bi bilo vredno modele, ki izkoriščajo naključje, na primer izjemno naključna drevesa [63], prevesti in preizkusiti na nalogi napovedovanja kvantilov (ali napovednih intervalov).

Napovedni intervali predstavljajo meje, znotraj katerih se z določeno gotovostjo nahajajo vrednosti odvisne spremenljivke, zato bi jih bilo načeloma možno uporabiti za odkrivanje osamelcev, ki ležijo daleč izven napovednih intervalov. Samo dolžino intervalov pa bi bilo možno razložiti z doprinosi posameznih atributov, kakor je bilo prikazano v klasifikaciji [64].

Klasična klasifikacija in regresija štejeta za osnovni paradigmi nadzorovanega učenja, nista pa edini. V nadaljnjem delu bi morali posvetiti pozornost še novejšim paradigmam, kot so večciljno učenje, delno nadzorovano učenje, ipd. Časovne vrste so že vrsto let zanimivo raziskovalno področje, v zadnjih letih pa se vse več pozornosti

posveča tako imenovanim podatkovnim tokovom. Pomanjkljivost množic, ki smo jih testirali, je, da je velika večina množic relativno majhnih za današnje razmere. Danes smo lahko celo preplavljeni s podatki, v obliki podatkovnih tokov. Bistvo obeh paradigem je, da novi podatki prihajajo v časovnih zaporedjih in pri slednji, da jih je veliko. Zanimivo je tudi, da se koncepti lahko spreminjajo s časom. Ocenjevanje zanesljivosti je pri obeh področjih zanimivo iz vseh že prej omenjenih razlogov, dodatno pa obstaja možnost odkrivanja sprememb konceptov. Da bi bile obravnavane metode uporabne na podatkovnih tokovih, jih bo potrebno optimizirati ali aproksimirati, saj je zgornja meja časovne zahtevnosti za metode, uporabne v tem kontekstu, linearna.

Omenjena področja oziroma raziskovalne teme so praktično še v povojih, vendar se v njih najverjetneje skrivajo novi odgovori na osnovna uporabniška vprašanja.

# *Praktični napotki za uporabo*

*A*

Za osnovni praktični primer imejmo neko podatkovno množico in vnaprej določen model, želimo pa izvesti analizo uporabnosti ocen zanesljivosti posameznih napovedi. Izbira modela lahko izvira iz predznanja, omejitev računskih sredstev ali časa, lahko pa iz nekih drugih razlogov. V tem primeru lahko analiziramo le različne pristope k ocenjevanju zanesljivosti posameznih napovedi. Če imamo opravka s klasifikacijskim problemom, potem imamo na voljo samo točkovne cenilke zanesljivosti. V nasprotnem primeru, če se spopadamo z regresijskim problemom, so na voljo tako točkovne kot intervalne cenilke. V obeh primerih je naslednji korak ponovno nekoliko odvisen od predznanja. Če obstaja argumentacija, zakaj določena metoda ocenjevanja zanesljivosti ustreza obravnavanemu primeru, potem je odločitev enolična. Vendar gre v večini primerov za ugibanje in je priporočljivo preizkusiti vse pristope, ne glede na predznanje.

Točkovne ocene zanesljivosti posameznih napovedi nudijo nekoliko omejeno stopnjo dodatne informacije, saj nam točkovna ocena zanesljivosti ene napovedi ne pove ničesar – koristne so šele v primerjavi dveh ali več napovedi med seboj. In ravno s primerjavo točkovnih ocen med seboj je možno ugotoviti, ali so točkovne ocene posamezne metode koristne. Za to je potrebna množica primerov, katerih odvisna spremenljivka je znana; za preprostost predpostavimo, da uporabljamo validacijsko množico. Za te primere najprej pridobimo napovedi modela, z njimi pa izračunamo napake napovedi. Za iste primere izračunamo tudi točkovne ocene zanesljivosti. Zanima nas, ali točkovne ocene zanesljivosti kaj razkrivajo o sami napaki, torej ali obstaja med njima neka odvisnost. Zanima nas Spearmanov rangirni koeficient korelaĺcije, saj obe vrednosti prihajata iz različnih porazdelitev in nimamo nobenega zagotovila, da bi obstajala linearna odvisnost. Zato vse vrednosti nadomestimo z njihovimi rangiranji, torej razvrstitvami v urejenih množicah. Pearsonov korelacijski koeficient rangov je iskani Spearmanov korelacijski koeficient. Ta nam pove, v kolikšni meri sta monotoni funkciji zanesljivosti in napake odvisni med seboj, torej koliko informacij nam ocene zanesljivosti nudijo o napaki posameznih napovedi. Metoda, ki doseže največjo stopnjo korelacije, je najbolj informativna za zadano kombinacijo podatkovne množice in modela.

Če se ukvarjamo s klasifikacijskim problemom, imamo na voljo še dva dodatna koraka analize koristnosti točkovnih ocen. Pri klasifikaciji lahko izračunamo referenčne ocene zanesljivosti kar iz posameznih napovedi modela (razdelek 4.2). Vprašanje je, ali lahko točkovne cenilke dajo boljše ocene zanesljivosti posameznih napovedi od mode-

lov samih. Odgovor na to vprašanje dobimo tako, da najprej za vse primere validacijske množice izračunamo referenčne ocene zanesljivosti in zanje izračunamo Spearmanov rangirni koeficient korelacije. Na koncu moramo preveriti, ali je kateri izmed korelacijskih koeficientov točkovnih cenilk značilno višji od korelacijskega koeficienta referenčne cenilke. V primeru, ko so koeficienti korelacije točkovnih cenilk neznačilno različni ali celo značilno nižji, nudi največ informacij o zanesljivosti posameznih napovedi kar model sam. Zadnje vprašanje pri analizi točkovnih ocen na klasifikacijskih problemih je, v kolikšni meri ocene zanesljivosti ločijo med pravilnimi in napačnimi napovedmi. Odgovor najlažje razberemo z grafičnim prikazom, kot so slike 4.2, 4.3 in 4.4.

Intervalno ocenjevanje je dobro definirano le nad zveznimi odvisnimi spremenljivkami. Ponovno poudarjamo, da je najbolje preizkusiti vse pristope, saj nam vsak izmed njih razkriva svoj pogled na podatke. Napovedni intervali so v praksi bolj informativni od točkovnih ocen, saj opisujejo intervale, znotraj katerih pričakujemo z vnaprej določeno gotovostjo prave vrednosti odvisne spremenljivke. O sami kvantizaciji zanesljivosti govorijo dolžine napovednih intervalov; lahko rečemo, da so napovedi s krajšimi intervali bolj zanesljive od napovedi z daljšimi intervali. Pri obnašanju napovednih intervalov obstajata dva ključna pojma: pravilnost obravnava doseganje zastavljene gotovosti vsebovanja odvisne spremenljivke in optimalnost, pri čemer so v povprečju krajši napovedni intervali bolj zaželeni. Številčna predstavitev teh dveh pojmov sta pokrivna verjetnost napovednih intervalov in relativni povprečni napovedni interval (predstavljena v razdelku 6.2), kjer za prvo vrednost želimo, da je čim bližje zastavljeni verjetnosti, za drugo pa želimo, da je čim manjša. Vendar kakor nakazujejo rezultati in prikazujejo slike v razdelku 6.3, same statistike ne zagotavljajo konformnosti napovednih intervalov. Da so napovedni intervali zares koristni, morajo biti konformni, torej se morajo čim bolje prilegati podatkom. V ta namen sta trenutno najboljši oziroma najprimernejši vizualizacijski tehniki, predstavljeni v razdelku 7.2.

Če sprostimo začetno omejitev določenosti modela, se raziskovalni prostor močno poveča. Po opravljenih analizah se lahko zgodi, da ne obstaja en najboljši model. Takrat se izplača preizkusiti pristope združevanja posameznih napovedi. Za točkovne cenilke je metodologijo moč najti v [45], za intervalne pa so postopki opisani v razdelku 7.3.

# LITERATURA


[1] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072.

[2] Igor Kononenko and Matjaž Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, 2007. ISBN 1904275214, 9781904275213.

[3] Zoran Bosnić and Igor Kononenko. Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering*, 67(3):504 – 516, 2008. ISSN 0169-023X. doi: 10.1016/j.datak.2008.08.001.

[4] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

[5] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[6] Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, 1896.

[7] Gene Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *J. Soc. Ind. Appl. Math., Ser. B, Numer. Anal.*, 2:205–224, 1965.

[8] C. Burges, B. Scholkopf, and A. Smola. *Advances in Kernel Methods: Support Vector Learning*. MIT press, Cambridge, MA, 1999.

[9] David E. Rumelhart and James L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Bradford Bks. MIT Press, 1986. ISBN 9780262631129.

[10] Brian D. Ripley and Nils L. Hjort. *Pattern Recognition and Neural Networks*. Cambridge University Press, New York, NY, USA, 1st edition, 1995. ISBN 0521460867.

[11] David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[12] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[13] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, July 1990. ISSN 0885-6125.

[14] Leo Breiman. Random forests. *Machine Learning*, 45 (1):5–32, 2001.

[15] Harris Drucker. Improving regressors using boosting techniques. In Douglas H. Fisher, editor, *ICML*, pages 107–115. Kaufmann, Morgan, 1997. ISBN 1-55860-486-3.

[16] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[17] Robert Tibshirani and Keith Knight. Model search and inference by bootstrap "bumping". Technical report, 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.8633.

[18] Pierre Geurts. Dual perturb and combine algorithm. In *Proceedings of AISTATS 2001, 8. International Workshop on Artificial Intelligence and Statistics*, pages 196–201, Key-West, Florida, January 2001.

[19] Kathleen M. Kerr and Gary A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*, 98: 8961–8965, 2000.

[20] Matthias Seeger. Learning with labeled and unlabeled data. Technical report, 2001. URL http://citeseer.uark.edu:8080/citeseerx/viewdoc/summary?doi=10.1.1.28.850.

[21] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39(1):1–38, 1977.

[22] Shumeet Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In Michael J. Kearns, Sara A. Solla, and David A. Cohn, editors, *NIPS*, pages 854–860. The MIT Press, 1998. ISBN 0-262-11245-0.

[23] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. pages 92–100. Morgan Kaufmann Publishers, 1998.

[24] Virginia R. de Sa. Learning classification with unlabeled data. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *NIPS*, pages 112–119. Morgan Kaufmann, 1993. ISBN 1-55860-322-0.

[25] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

[26] Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann, 1998.

[27] Craig Saunders, Alex Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 722–726, 1999.

[28] Matjaž Kukar. *Ocenjevanje zanesljivosti klasifikacij in cenovno občutljivo kombiniranje metod strojnega učenja*. PhD thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2001.

[29] Zoran Bosnić, Igor Kononenko, Marko Robnik-Šikonja, and Matjaž Kukar. Evaluation of prediction reliability in regression using the transduction principle. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, volume 2, pages 99–103 vol.2, 2003.

[30] Lucia Breierova and Mark Choudhari. *An Introduction to Sensitivity Analysis*. Massachusetts Institute of Technology. System Dynamic in Education Project, Cambridge, 1996.

[31] Sherif Hashem. Sensitivity analysis for feedforward artificial neural networks with differentiable activation functions. In *Proceedings of the 1992 International Joint Conferences on Neural Networks*, pages 419–424. IEEE Press, 1992.

[32] Linda C. van der Gaag and Uffe Kjaerulff. Making sensitivity analysis computationally efficient. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 317–325. Morgan Kaufmann Publishers, 2000.

[33] Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *NIPS*, pages 196–202. MIT Press, 2000.

[34] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526, 2002.

[35] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236: 333–380, 1937.

[36] David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks, Orlando, FL (IEEE–ICNN'94)*, pages 55–60. IEEE-Press, 1994.

[37] Robert Tibshirani. A comparison of some error estimates for neural network models. *Neural Comput.*, 8(1):152–163, January 1996. ISSN 0899-7667. doi: 10.1162/neco.1996.8.1.152.

[38] Tom Heskes. Practical confidence and prediction intervals. In Michael Mozer, Michael I. Jordan, and Thomas Petsche, editors, *NIPS*, pages 176–182. MIT Press, 1996.

[39] Achilleas Zapranis and Efstratios Livanis. Prediction intervals for neural network models. In *Proceedings of the 9th WSEAS International Conference on Computers*, ICCOMP'05, pages 76:1–76:7, Stevens Point, Wisconsin, USA, 2005. World Scientific and Engineering Academy and Society (WSEAS). ISBN 960-8457-29-7.

[40] Durga L. Shrestha and Dimitri P. Solomatine. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2): 225 – 235, 2006. ISSN 0893-6080.

[41] Sejong Oh. A new dataset evaluation method based on category overlap. *Comp. in Bio. and Med.*, 41(2): 115–122, 2011.

[42] Roger Koenker and George Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[43] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, pages 101–474, 2002.

[44] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, December 2006. ISSN 1532-4435.

[45] Zoran Bosnić. *Ocenjevanje zanesljivosti posameznih napovedi z analizo občutljivosti regresijskih modelov*. PhD thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2007.

[46] Darko Pevec, Zoran Bosnić, and Igor Kononenko. Comparison of approaches for estimating reliability of individual classification predictions. In Marko ... [et al.] Bohanec, editor, *IS2009*, pages 46–49. Ljubljana: Institut Jožef Stefan, 2009.

[47] Darko Pevec, Zoran Bosnić, and Igor Kononenko. Individual prediction reliability estimates in classification and regression. In José María Martínez-Martínez Joan Vila-Francés Rafael Magdalena-Benedito, Marcelino Martínez-Sober and Pablo Escandell-Montero, editors, *Intelligent Data Analysis for Real-Life Applications: Theory and Practice*, pages 35–56. Hershey: IGI Global, 2012.

[48] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 97–104, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

[49] Darko Pevec. Ocenjevanje zanesljivosti posameznih klasifikacij z lokalnimi metodami. Diplomsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Ljubljana, 2009.

[50] Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[51] Stephen Portnoy and Roger Koenker. The Gaussian Hare and the Laplacian Tortoise: Computability of Squared- Error versus Absolute-Error Estimators. *Statistical Science*, 12(4):279–296, 1997.

[52] Darko Pevec, Erik Štrumbelj, and Igor Kononenko. Evaluating reliability of single classifications of neural networks. In Andrej Dobnikar, Uroš Lotrič, and Branko Šter, editors, *11th International Conference on Adaptive and Natural Computing Algorithms (1)*, volume 6594 of *Lecture Notes in Computer Science*, pages 22–30. Springer Berlin Heidelberg, 2011.

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

[54] John A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979. ISSN 00359254.

[55] Darko Pevec and Igor Kononenko. Input dependent prediction intervals for supervised regression. *Intelligent Data Analysis*, 18(5), 2014, v tisku.

[56] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.

[57] Belur V. Dasarathy and Belur V. Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, 1979. ISSN 0018-9219. doi: 10.1109/PROC.1979.11321.

[58] Darko Pevec and Igor Kononenko. Model selection with combining valid and optimal prediction intervals. pages 653–658, Los Alamitos, CA, USA, 2012. IEEE Computer Society. ISBN 978-1-4673-5164-5.

[59] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.*, 53(1-2):23–69, October 2003. ISSN 0885-6125.

[60] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primož Škraba. Persistence-based clustering in riemannian manifolds. In *Proc. 27th Annu. ACM Sympos. on Comput. Geom.*, pages 97–106, June 2011.

[61] Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9:371–421, March 2008.

[62] Lynne Billard and Edwin Diday. From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487, June 2003.

[63] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1): 3–42, 2006. ISSN 0885-6125.

[64] Erik Štrumbelj, Igor Kononenko, and Marko Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10):886–904, October 2009. ISSN 0169-023X.

# SLOVARČEK IZRAZOV

*residual* (residual)
razlika med zabeleženo oziroma opazovano vrednostjo in oceno vrednosti naključne spremenljivke

*kvantil* (quantile)
točka, ki deli kumulativno funkcijo verjetja na enake dele; znane podpomenke so kvartil (deli distribucijo na štiri), decil (na deset), percentil (na sto)

*homoskedastičnost* (homoscedasticity)
statistični izraz, ki označuje homogenost, torej konstantnost in končnost variance naključne spremenljivke

*heteroskedastičnost* (heteroscedasticity)
izraz, ki je komplement prejšnjega; pomeni, da varianca naključne spremenljivke ni homogena, torej je v različnih podpopulacijah različna, ali celo funkcijsko odvisna

*stremljenje* (bootstraping)
statistični pristop ponavljanja vzorčenja z vračanjem, ki omogoča ocenjevanje vzorčne porazdelitve poljubne statistike (kot je varianca)

*maksimalno verjetje* (maximum likelihood)
gre za pristop ocenjevanja parametrov statističnih modelov na osnovi zbranega vzorca tako, da se predpostavi, da je dobljeni vzorec najbolj verjeten, zanimajo nas pa parametri, ki so ustvarili dobljeni vzorec

*točkovna ocena zanesljivosti* (point-wise reliability estimate)
ocena zanesljivosti (posamezne napovedi), ki ima le količinsko interpretacijo, v smislu, da je možna le primerjava dveh ali več ocen med seboj (večje, manjše ali enako)

*interval zaupanja* (confidence interval)
interval, v katerem se z dano gotovostjo nahaja ocenjevani parameter

*intervalna ocena zanesljivosti — napovedni interval*
(interval reliability estimate — prediction interval)
ocena zanesljivosti v obliki intervala, v katerem se z dano gotovostjo nahaja prava vrednost naključne spremenljivke

*konformnost napovednih vrednosti* (prediction conformity)
s tem izrazom označujemo pravilno prileganje napovedanih vrednosti pogojni porazdelitvi naključne spremenljivke