Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tomaž Kariž

Ocenjevanje zanesljivosti napovedi skupinskih modelov

magistrsko delo

študijski program druge stopnje
računalništvo in informatika

Mentor: izr. prof. dr. Janez Demšar

Ljubljana, 2015

Izjava o avtorstvu magistrskega dela

Spodaj podpisani Tomaž Kariž sem avtor magistrskega dela z naslovom:

Ocenjevanje zanesljivosti napovedi skupinskih modelov

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Janeza Demšarja,
- so elektronska oblika magistrskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela,
- soglašam z javno objavo elektronske oblike magistrskega dela v zbirki "Dela FRI".

V Ljubljani, 25. junija 2015 Podpis avtorja:

Zahvaljujem se mentorju izr. prof. dr. Janezu Demšarju za nasvete, razlage in spodbudo pri izdelavi magistrskega dela. Posebna zahvala gre tudi bratu in punci za podporo in motivacijo ter staršema za finančno in moralno podporo skozi vsa leta študija.

Kazalo

1 Uvod **1**

2 Pregled obstoječega dela **3**

2.1 Metode specifične za model 3

2.2 Metode neodvisne od modela 10

3 Ocenjevanje zanesljivosti skupinskih modelov **17**

3.1 Regresijske metode 17

3.1.1 Varianca notranjih modelov (VAR) 17

3.1.2 K najbližjih stanj odstopanj (KNSO) 18

3.1.3 Stacking notranjih modelov (SNM) 20

3.2 Klasifikacijske metode 22

3.2.1 Delež izglasovanega razreda (Ratio) 22

3.2.2 Klasifikacijska preciznost (Prec) 22

3.2.3 KNN notranjih modelov (KnnNM) 25

4 Postopek testiranja metod **29**

4.1 Podatkovne množice 30

4.2 Mere statistike 31

4.3 Uporabljeni skupinski modeli 33

4.4 Primerjava uspešnosti metod 34

5 Rezultati **37**

5.1 Regresijske podatkovne množice 37

5.1.1 Naključni gozdovi 38

5.1.2 Bagging 41

5.2 Klasifikacijske podatkovne množice 44
5.2.1 Naključni gozdovi 45
5.2.2 Bagging 48

6 Sklepne ugotovitve **53**

Seznam uporabljenih kratic

kratica	celoten zapis
BAGV	Varianca modela bagging
BVCK	Kombinacija BAGV in CNK
CNK	Lokalno modeliranje napovedne napake
DENS	Ocenjevanje gostote
KnnNM	K najbližjih sosedov notranjih modelov
KNSO	K najbližjih stanj odstopanj
LCV	Lokalno prečno preverjanje
MHN	Mahalanobisova razdalja
MHNC	Mahalanobisova razdalja do središča
ORef	Referenčna ocena
Prec	Klasifikacijska preciznost
Ratio	Delež izglasovanega razreda
SAb	Analiza občutljivosti (bias)
SAv	Analiza občutljivosti (variance)
SNM	Stacking notranjih modelov
VAR	Varianca notranjih modelov

Povzetek

V današnjem svetu je zanesljivost napovedovanja zelo pomembna, predvsem na področjih, kot sta recimo zdravstvo in finance, kjer ne bi radi napovedali česa, v kar nismo dovolj prepričani. V strojnem učenju se za reševanje teh problemov raziskuje metode, ki bi nam skušale oceniti, kako zanesljive so naše napovedi. Pri ocenjevanju zanesljivosti napovedi obstajata dve vrsti metod: takšne, ki se specializirajo za točno določen model in takšne, ki ne predpostavljajo vnaprej vrste modela. Prve lahko upoštevajo dodatne informacije pri določanju zanesljivosti, saj lahko uporabijo parametre, ki so specifični za model, kot dodatno informacijo. Druge pa imajo to lastnost, da delujejo na vseh modelih. V delu predstavimo nekaj novih metod, ki delujejo na skupinskih modelih, torej spadajo med tiste, ki so specifične za določen model. Metode delujejo tako na klasifikacijskih kot tudi na regresijskih podatkovnih množicah. Uspešnost metod ovrednotimo s Pearsonovim korelacijskim koeficientom v primeru regresijskih problemov in Wilcoxon-Mann-Whitneyevo statistiko v primeru klasifikacijskih. Razvite metode primerjamo z že obstoječimi in rezultate prikažemo z grafom rangov kritične razdalje.

Ključne besede: strojno učenje, ocenjevanje zanesljivosti, zanesljivost napovedi, skupinski modeli

Abstract

In today's world, the reliability of a prediction is very important, especially in areas such as health and finance, where we do not want to make predictions that are not sufficiently reliable. To solve these problems in the context of machine learning, methods are being researched that assess the reliability of predictions. There are two types of methods: those specialized for a specific model and those who do not presume in advance the model type. The first may take into account additional information in determining the reliability, because they can use the parameters that are specific to the model as additional information. Others, however, are applicable to all models. In this work, we present some methods that operate on ensemble models, therefore, they are among those that are specific to a particular model. Methods operate on both the classification as well as regression datasets. Performance of methods is evaluated by Pearson correlation coefficient in the case of regression problems and Wilcoxon-Mann-Whitney statistics in the case of classification. The developed methods are compared with existing ones. We also show the results using critical distance diagrams.

Keywords: machine learning, reliability assessment, prediction reliability, ensemble models

Poglavje 1

Uvod

Strojno učenje se ukvarja z napovedovanjem ciljne spremenljivke. Množice, ki predstavljajo podatke s pomočjo katerih želimo napovedati vrednost ciljne spremenljivke, so lahko pridobljene iz različnih področij. Pogosto se želimo prepričati o zanesljivosti napovedi, saj napoved modela ni vedno pravilna oziroma dovolj blizu pravilnega, če govorimo o zveznih spremenljivkah. Kot primer lahko vzamemo napovedovanje bolezni pacienta, kjer je zanesljivost napovedi pomemben podatek, saj je napačna klasifikacija lahko življenjskega pomena. Ocenjevanje zanesljivosti ni enostaven problem, zato je bilo razvitih že precej metod, ki skušajo karseda dobro oceniti zanesljivost napovedi.

V svojem delu bomo predstavili nekaj novih pristopov k reševanju problema ocenjevanja zanesljivosti napovedi. Metode, ki jih bomo predstavili, so namenjene specifičnim oblikam modelov, specifično skupinskim modelom. Nobena izmed metod ne predpostavi vrste notranjih modelov, torej je vseeno ali imamo notranje modele različnih vrst (stacking) ali pa vse modele iste vrste (na primer bagging). Metode torej notranje modele obravnavajo kot črne škatle, kjer vsaka poda svojo napoved za podan primer. Pri dodani informaciji o modelu in z dostopom do notranjih modelov nas torej zanima, ali lahko s predpostavko, da model spada med skupinske modele in so nam te informacije na voljo, dosežemo boljše ocene zanesljivosti, kot s splošnejšimi metodami. Večina razvitih metod deluje tako na regresijskih kot tudi na klasifikacijskih podatkovnih množicah. Za ovrednotenje uspešnosti metod smo uporabili znane pristope in sicer izračun korelacije med napako in oceno zanesljivosti pri regresijskih problemih in Wilcoxon-Mann-

Whitneyevo U statistiko pri klasifikacijskih. Razvite metode smo nato primerjali z že obstoječimi metodami, ki smo jih opisali v 2. poglavju.

Poglavje 2 je namenjeno pregledu obstoječega dela. Tu opišemo nekatere obstoječe metode, ki so specifične za določen model in tiste, ki so neodvisne od modela. V tretjem poglavju predstavimo razvite metode, ki delujejo na skupinskih modelih. Opišemo tri metode, ki delujejo na regresijskih problemih in tri, ki so namenjene reševanju klasifikacijskih problemov, vendar predstavimo tudi način, kako metode namenjene ocenjevanju zanesljivosti napovedi na regresijskih problemih prilagodimo za delovanje na klasifikacijskih problemih in obratno. V četrtem poglavju opišemo testiranje metod. Tu najprej opišemo podatke, na katerih smo metode testirali. Sledi opis statističnih mer, s katerimi smo metode primerjali ter predstavitev uporabljenih skupinskih modelov. Na koncu poglavja predstavimo tudi Graf rangov s kritično razdaljo, ki nam prikaže, kako so se mere odrezale. V petem poglavju predstavimo rezultate testiranj tako na regresijskih kot tudi na klasifikacijskih problemih. Vsak skupinski model smo testirali s 50 in 100 notranjimi modeli in prikazali uspešnost metod pri ocenjevanju zanesljivosti njegove napovedi. Metode smo pri vsakem skupinskem modelu rangirali glede na uspešnost na vseh učnih množicah od najuspešnejše do najmanj uspešne. V zadnjem poglavju predstavimo ugotovitve testiranj in predloge za nadaljnje delo.

Poglavje 2

Pregled obstoječega dela

Za reševanje problema ocenjevanja zanesljivosti je bilo razvito precej metod, ki jih lahko razdelimo na dve vrsti: metode, ki so specifične za nek model in metode, ki so neodvisne od modela. Prednost prvih je, da lahko izkoristijo dodatne informacije, ki so jim na voljo zaradi poznavanja strukture modela. Te dodatne informacije lahko pripomorejo k pravilnejši oceni zanesljivosti. Slabost teh metod je, da delujejo le na specifičnem modelu, zato so v primeru drugačnega modela neuporabne. Tu vidimo prednost drugih, saj zaradi predpostavke, da o modelu ne vedo nič - včasih rečemo, da model obravnavajo kot črno škatlo - delujejo na poljubnem modelu. Slednje metode torej ne uporabijo dodatne informacije, ki jih ponuja model in posledično postanejo neodvisne od modela.

Poleg tega so ocene zanesljivosti lahko različno predstavljene. Običajno je ocena zanesljivosti podana v obliki intervala ali pa v obliki numerične vrednosti. Metode, ki smo razvili in tiste, s katerimi jih bomo primerjali, podajo oceno zanesljivosti v numerični obliki. V nadaljevanju bomo opisali nekaj metod, ki so bile razvite za specifičen model in nekaj takih, ki so neodvisne od modela.

2.1 Metode specifične za model

V tem razdelku opišemo nekaj pristopov k ocenjevanju zanesljivosti napovedi specifičnega modela, ki so bili uporabljeni v raziskavah. Najprej bomo prikazali zelo enostavne pristope ocenjevanja zanesljivosti, ki so bili uporabljeni za izboljšanje

napovedi skupinskega modela. Sledili bodo pristopi specifični za model podpornih vektorjev, nevronskih mrež in skupinskih modelov, ki imajo kot notranji model drevo.

Izboljšanje napovedi skupinskih modelov

Monteith in Martinez [10] naštevajo preproste načine ocenjevanja zanesljivosti posameznega primera klasifikatorjev odločitveno drevo, naivni Bayes, K-najbližjih sosedov in nevronske mreže. Nekatere izmed mer ocene zanesljivosti napovedi za posamezen klasifikator so prikazane spodaj.

Odločitveno drevo:

- število primerov v listu, v katerega sodi opazovani primer,
- odstotek pravilno napovedanih primerov pri preverjanju izpusti enega na primerih, ki so bili uvrščeni v isti list in
- višina, na kateri se nahaja list, s katerim napovemo razred.

K-najbližjih sosedov:

- odstotek sosedov, ki imajo razred enak večinskemu razredu sosedov pri iskanju sosedov,
- odstotek pravilno klasificiranih sosedov pri preverjanju izpusti enega na učni množici in
- podobno kot prejšnja, le da meri odstotek samo od sosedov, ki imajo večinski razred med sosedi.

Naivni Bayes:

- verjetnost razreda, ki je napovedan,
- razdalja med verjetnostjo napovedanega razreda in verjetnostjo drugega najbolj verjetnega razreda in
- razdalja med verjetnostjo napovedanega razreda in vsoto verjetnosti ostalih razredov.

Nevronske mreže:

- aktivacijski izhod izbrane klasifikacije,

- odstotek pravilno napovedanih sosedov pri preverjanju izpusti enega na učni množici, pri čemer razdalja predstavlja razdaljo med aktivacijskimi izhodi in
- razlika med dvema najvišjima aktivacijskima izhodoma.

Avtorji želijo v delu izboljšati napoved modela s pomočjo omenjenih mer, da bi prišli do boljše napovedi skupinskega modela, ki so ga sestavljali klasifikatorji opisani zgoraj. To dosežejo tako, da za vsak klasifikator s prečnim preverjanjem na učni množici izmerijo korelacijo med pravilnostjo napovedi in oceno zanesljivosti napovedi za vsako posamezno mero. Za neznani primer x_i so za posamezen klasifikator ocenili zanesljivost napovedi z merami definirami na tem klasifikatorju in te vrednosti normalizirali glede na minimalno in maksimalno zabeleženo vrednost posamezne mere na učni množici. S kombinacijo mer so pridobili skupno mero zanesljivosti za posamezen klasifikator, ki se je izkazala za boljšo od posameznih mer, kar je veljalo pri vseh klasifikatorjih. Skupno mero so definirali kot skalarni produkt vektorja normaliziranih ocen zanesljivosti primera x_i in vektorja normaliziranih korelacij. Poleg izboljšane ocene zanesljivosti napovedi so z njo uspeli tudi izboljšati napoved skupinskega modela.

Zanesljivost napovedi metode podpornih vektorjev

Zanesljivost napovedi so ocenili tudi pri metodi podpornih vektorjev. Saunders in dr. [13] opišejo postopek s katerim ne le podajo ocene zanesljivosti napovedi, temveč tudi oceno kredibilnosti, ki je mera zanesljivosti podatkov, na podlagi katerih smo podali napoved. Pristop je namenjen ocenjevanju zanesljivosti napovedi v binarnih klasifikacijskih problemih. Metoda podpornih vektorjev temelji na optimizaciji postavitve hiperravnine za ločitev vrednosti razreda. Vsak izmed vektorjev v prostoru ima po rešeni optimizacijski funkciji svoj Lagrangeov multiplikator, ki ga bomo v nadaljevanju označili z α_i, kjer bo i predstavljal indeks vektorja, kateremu multiplikator pripada. Izkaže se, da so ti multiplikatorji v večini nizki oz. enaki 0 in je le nekaj takih vektorjev, ki imajo visoko vrednost. Ti vektorji določajo hiperravnino in jim zato pravimo podporni vektorji. Pri ocenjevanju zanesljivosti klasifikacije novega primera x_{n+1} Saunders in dr. [13] optimizacijski problem rešijo dvakrat in sicer enkrat, ko je nov primer v razredu -1 in enkrat, ko

je v +1. Za vsako izmed dobljenih hiperravnin so podali oceno, kako nenavaden je nov primer pri tej hiperravnini. Ideja *nenavadnosti* primera je v tem, da je precej majhna verjetnost, da bo dani primer igral veliko vlogo pri podpiranju hiperravnine, saj je podpornih vektorjev zelo malo. Denimo, da imamo učno množico z novim primerom x_{n+1}

$$ucna = \{x_1, x_2, ...x_n, x_{n+1}\}$$

Po maksimiziranju optimizacijske funkcije pridobimo $n+1$ Lagrangeovih multiplikatorjev $\alpha_1, ...\alpha_{n+1}$. Če bi nov primer x_{n+1} imel največjo α vrednost, bi bilo to "sumljivo", saj je verjetnost, da pride do tega, zelo majhna in sicer $\frac{1}{n+1}$. Nenavadnost primera je definirana z enačbo

$$P\_vrednost = \frac{nr : \alpha_i \geq \alpha_{n+1}}{n+1}$$

Na podlagi ocene nenavadnosti na obeh hiperravninah se njihov model odloči za napoved. Če je p-vrednost visoka, je bil vektor najbrž precej nepomemben, torej na območju, ki je dobro definirano in ni igral velike vloge, v kolikor pa je nizka, je bolj verjetno, da je bil primer napačno klasificiran. Napoved modela je torej tisti razred, kjer je bila P-vrednost večja. Zanesljivost klasifikacije je

$$zanesljivost = 1 - P_2,$$

kjer P_2 predstavlja oceno p-vrednosti primera, ki ni bil napovedan. Če je x_{n+1} v obeh primerih deloval kot podporni vektor, je ta vrednost nižja. Poleg zanesljivosti pa dobimo informacijo o kredibilnosti, ki nam pove o kvaliteti podatkov.

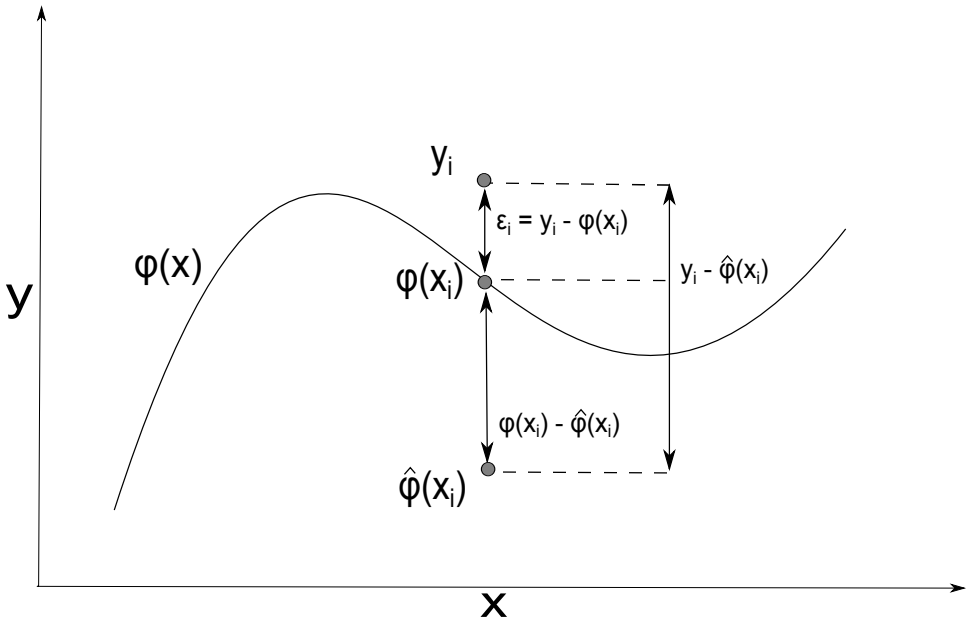
$$kredibilnost = P_1.$$

Ker je mera zanesljivosti vplivala na napoved modela, so avtorji izvedli tudi primerjavo s standardnim modelom. Z rezultati testiranj so pokazali, da je model še vedno primerljiv z osnovno implementacijo in je pri napačnih klasifikacijah podal nizko oceno kredibilnosti.

Zanesljivost napovedi nevronskih mrež

Precej raziskav na primer [18] se je osredotočilo na ocenjevanje zanesljivosti napovedi nevronskih mrež. Varianco σ_p^2, ki predstavlja šum pri napovedi, lahko definiramo kot vsoto dveh neodvisnih varianc šumov σ_m^2 in σ_ε^2. Prva (σ_m^2) predstavlja

šum modela, saj je ta lahko pristranski in neoptimalen. Druga (σ_ε^2) pa predstavlja šum podatkov, do katerega pride zaradi nepopolnih ali napačnih podatkov znotraj naših podatkovnih množic. Če poznamo σ_m^2 za primer x_i, ki ga napovedujemo, lahko določimo interval zaupanja, ki predstavlja interval, na katerem se z določeno verjetnostjo nahaja prava vrednost funkcije $\varphi(x_i)$, ki nam ni znana. Interval zaupanja torej opisuje distribucijo $\varphi(x_i) - \hat{\varphi}(x_i)$. Skrita funkcija φ slika x_i v njegovo pravo vrednost, ki se od opazovane vrednosti y_i razlikuje v tem, da ne vsebuje šuma. Bolj zanimivi od intervalov zaupanja so napovedni intervali, ki za posamezen primer predstavljajo interval, znotraj katerega se z določeno verjetnostjo nahaja opazovana vrednost y_i, torej opisujejo distribucijo $y_i - \hat{\varphi}(x_i)$. Na sliki 2.1 ordinatna os predstavlja vrednost ciljne spremenljivke, abscisna pa primere, ki jih ocenjujemo. Iz nje je razvidno, da je interval zaupanja vsebovan znotraj intervala napovedi.



Slika 2.1: Prikaz razmerja med napovedjo modela $\hat{\varphi}(x_i)$, opazovano vrednostjo y_i, pravo vrednostjo primera $\varphi(x_i)$ in komponento šuma ϵ_i.

Oceno variance šuma modela lahko dobimo s pomočjo metode stremljenja. Ta

metoda generira M učnih množic, pri čemer posamezno učno množico pridobi tako, da iz celotne učne množice velikosti n vzame n primerov s ponavljanjem. Stremljenje je uporabljeno tudi v modelu bagging, ki je opisan v poglavju 4. Stremljenje je namenjeno predvsem izboljšanju napovedi modela, predvsem pri modelih, ki imajo visoko varianco napovedi. Ocena variance šuma modela je definirana kot varianca napovedi notranjih modelov.

$$\sigma_m^2(x) = \frac{\sum_{i=1}^{M} (C_i(x) - \overline{C(x)})^2}{M}$$

Za napovedne intervale potrebujemo tudi varianco šuma podatkov, ki jo lahko ocenimo z metodo maksimalnega verjetja. Pri tem predpostavimo, da nam nevronska mreža g, ki je naučena na učni množici, da dober približek skrite funkcije φ. Ob tej predpostavki lahko oceno variance šuma podatkov pridobimo tako, da ustvarimo nov model nevronskih mrež f, ki ima za ciljno spremenljivko kvadratno napako napovedi nevronske mreže g. Zaradi tega je optimizacijska funkcija novega modela f, ki jo minimiziramo, definirana kot

$$\sum_{i=1}^{n} ((g(x_i) - y_i)^2 - f(x_i))^2,$$

kjer n predstavlja velikost učne množice, x_i pa trenuten primer, s pomočjo katerega učimo model f. S tem je napoved modela f približek variance šuma podatkov za dani primer ($\hat{\sigma}_\varepsilon^2(x_i) \approx f(x_i)$). Nix in Weigend predlagata rešitev problema tako, da imamo namesto dveh nevronskih mrež le eno, ki ima dva izhoda in sicer enega za napoved in enega za oceno šuma podatkov [11].

Napovedni interval za primer x_i ima sredino v povprečju napovedi notranjih modelov in je definiran kot:

$$g_{avg}(x_i) \pm Z_{\frac{\alpha}{2}} \hat{\sigma}_p^2(x_i)$$

Avtorja sta merila uspešnost napovednega intervala z mero pokrivne verjetnosti napovednega intervala, ki meri ali je v določenem odstotku napovedana vrednost znotraj intervala. Kombinacija opisanih metod se je izkazala za uspešno saj je 95% napovedni interval vseboval vrednost v 95.9% primerih.

Zanesljivost napovedi naključnih gozdov

Bhattacharyya [1] poda oceno zanesljivosti naključnih gozdov s pomočjo konformnega napovednega ogrodja. Konformno napovedno ogrodje [14] je eden izmed novejših pristopov k ocenjevanju zanesljivosti pri klasifikacijskih problemih, ki je bilo delno uporabljeno v [13] na metodi podpornih vektorjev, čeprav ga takrat še niso tako imenovali. Pri dani učni množici $(x_1, y_1), ...(x_n, y_n)$, kjer x_i predstavlja vektor atributov in y_i njegovo opazovano vrednost nam ogrodje omogoča, poleg same napovedi, ob poljubno definiranem ϵ podati tudi množico razredov τ za katero velja, da je pravi razred element te množice z verjetnostjo vsaj $1 - \epsilon$. Ideja ogrodja je, da pri napovedi novega primera x_{n+1} poizkusimo vsako možno vrednost razredne spremenljivke in izmerimo atipičnost primera pri izbranem razredu. Obstajajo različne mere ocenjevanja atipičnosti primera relativno na druge primere, vendar so vse oblike:

$$F(\{(x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ..., (x_{n+1}, y_{n+1})\}, (x_i, y_i)) \rightarrow \alpha_i,$$

kjer α_i predstavlja mero atipičnosti primera (x_i, y_i) v dani množici. Po izračunu atipičnosti vseh primerov množice lahko izračunamo p-vrednost

$$P((x_1, y_1), ...(x_{n+1}, y_{n+1})) = \frac{\#i : \alpha_i \geq \alpha_{n+1}}{n+1}$$

torej preštejemo, koliko atipičen je bil v primerjavi z ostalimi. Napoved modela je razred, pri katerem je p-vrednost največja. Posamezen razred je vsebovan v množici τ natanko takrat, ko je p-vrednost pri tem razredu večja od ϵ. Zanesljivost je definirana kot $1 - P_2$, kjer P_2 predstavlja drugo največjo p-vrednost. Poleg zanesljivosti je mogoče oceniti tudi kredibilnost, ki je kar p-vrednost napovedanega razreda. Bhattacharyya [1] predstavi tri mere atipičnosti primera, ki so opisane v nadaljevanju. Dve izmed mer uporabljata bližino dveh primerov, ki je definirana kot delež dreves, v katerih se oba primera nahajata v istem listu.

1.) Delež v naključnih gozdovih

Mera atipičnosti primera je definirana kot $1 - d$, kjer d predstavlja delež dreves, ki pravilno napovedo primer. P-vrednost se izračuna kot

$$Pvalue(new, c) = \frac{\#i : \alpha_i^c \geq \alpha_{new}^c \ \wedge \ y_i = c}{|F|},$$

kjer $|F|$ predstavlja množico primerov iz učne množice, ki imajo razred enak c, skupaj z novim primerom. α_i^c predstavlja mero atipičnosti i-tega primera, ki ima vrednost razredne spremenljivke c. P-vrednost razreda c nam pove kakšen je delež primerov z razredom c, ki so jih drevesa slabše napovedala v primerjavi z napovedjo novega primera, če predpostavimo, da je pravilen razred primera c.

2.) Bližina relativne okolice

Huazhen in dr. [17] so mero definirali kot

$$\frac{povprečna\ bližina\ K-najbližjih\ sosedov\ iz\ množice\ A}{povprečna\ bližina\ K-najbližjih\ sosedov\ iz\ množice\ B},$$

kjer množica B vsebuje primere z enako vrednostjo razredne spremenljivke kot primer, ki ga ocenjujemo in A vsebuje vse ostale primere. Bhattacharyya [1] mero izboljša s tem, da povprečno bližino K-najbližjih sosedov razreda c normalizira tako, da jo deli s povprečno bližino K-najbližjih sosedov vseh primerov razreda c.

3.) Število bližnjih primerov

Mera določi prag $\theta \in [0, 1]$. Za vse primere, ki imajo bližino večjo od θ velja, da so blizu primera. Mera atipičnosti danega primera je definirana kot

$$\alpha_{primer}^c = 1 - (\#i : bližina(x_i, primer) > \theta \wedge y_i = c).$$

P-vrednost se izračuna enako, kot pri meri atipičnosti s pomočjo deleža v naključnih gozdovih.

Mere niso preveč vplivale na samo napoved modela, saj je ta ostala primerljiva z osnovnim modelom. Poleg tega so vse tri metode na večini podatkovnih množic dosegle velik delež primerov, kjer je množica τ vsebovala le en razred.

2.2 Metode neodvisne od modela

V tem delu opišemo nekaj metod, ki so neodvisne od modela. Nekatere izmed metod - tiste, ki imajo poleg imena navedeno tudi kratico - smo uporabili pri testiranju za primerjavo z metodami razvitimi v tem delu.

Mahalanobisova razdalja (MHN)

Metoda uporabi za račun razdalje med dvema primeroma Mahalanobisovo razdaljo, ki je definirana kot

$$MHN(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)},$$

kjer x in y predstavljata vektorja, med katerima želimo izmeriti razdaljo, in S predstavlja kovariančno matriko. Razlika med Evklidsko in Mahalanobisovo razdaljo je v tem, da slednja upošteva razliko varianc v posameznih dimenzijah in kovarianco med spremenljivkami. Evklidska razdalja ima smisel takrat, ko so enote v dimenzijah atributov enake in so atributi neodvisni med seboj. Metoda poda oceno zanesljivosti tako, da najprej poišče K najbližjih sosedov primera, za katerega želimo podati oceno zanesljivosti in poda oceno kot vsoto razdalj do sosedov. Večja vsota predstavlja manjšo zanesljivost, torej pričakujemo pozitivno korelacijo z napako napovedi. V naših testiranjih smo poiskali 3 najbližje sosede.

Mahalanobisova razdalja do središča (MHNC)

Podobno kot zgoraj opisana metoda tudi ta poda oceno zanesljivosti glede na Mahalanobisovo razdaljo ampak ne meri razdalje do sosedov okoli primera, ki smo ga napovedali, temveč do središčne točke v prostoru. Prostor sestavljajo vsi primeri, ki jih imamo na voljo, središče tega pa je vektor, ki vsebuje povprečne vrednosti vsake dimenzije. Metoda torej zanesljivost poda v odvisnosti od razdalje do središča prostora.

Analiza občutljivosti (SAv, SAb)

Analiza občutljivosti [3] meri občutljivost napovedi modela pri vstavljanju dodatnih primerov v učno množico. Primer, ki ga ocenjujemo (v nadaljevanju označen z p) najprej napovemo in dobimo napoved K. Definiramo množico E, ki bo predstavljala intenziteto spreminjanja napovedi, na primer $E = [0.01, 0.1, 0.5, 1.0, 2.0]$. Za vsak $\alpha \in E$ ustvarimo novo učno množico, ki se od začetne učne množice razlikuje le v tem, da ima dodan nov primer z enakimi atributi kot p, vendar ima vrednost razredne spremenljivke $K + \alpha d$. Vrednost d je razlika med najvišjo in

najnižjo vrednostjo razredne spremenljivko v učni množici. Nato ustvarimo nov model, ki ga naučimo na tej učni množici in z njim dobimo napoved K_α. Na enak način dobimo napoved $K_{-\alpha}$ tako, da za vrednost razredne spremenljivke zapišemo $K - \alpha d$. Oceno zanesljivosti definiramo lahko definiramo na dva načina.

$$SA_v = \frac{\sum_{\alpha \in E}(K_\alpha - K_{-\alpha})}{|E|} \tag{2.1}$$

$$SA_b = \frac{\sum_{\alpha \in E}(K_\alpha - K) + (K_{-\alpha} - K)}{2|E|} \tag{2.2}$$

V testiranje smo vključili obe možnosti. Pri SA_b smo upoštevali absolutno vrednost ocene zanesljivosti.

Varianca modela bagging (BAGV)

Bagging je model, ki ga uvrščamo med skupinske modele. Notranje modele naučimo na različnih učnih množicah, ki so ustvarjene tako, da iz originalne učne množice velikosti N vzamemo N primerov s ponavljanjem. Bagging kot napoved vrne povprečno vrednost napovedi notranjih modelov, ki jo bomo označili s K. Bagging variance poda oceno zanesljivosti kot varianco napovedi notranjih modelov (2.3), ki nam pove, kako enotni so si notranji modeli.

$$BAGV = \frac{\sum_{i=1}^{k}(K_i - K)^2}{k} \tag{2.3}$$

Na podatkovnih množicah z diskretno ciljno spremenljivko je ocena zanesljivosti definirana kot

$$XS = [dist(P(C_1), P(S))+, ..., +dist(P(C_n), P(S))]$$

$$K = \frac{\sum_i^n XS_i}{n}$$

$$BAGV = \frac{\sum_i^n (XS_i - K)^2}{n}$$

$$ocena\_zanesljivosti = 1 - BAGV,$$

kjer $dist(P(C_i), P(S))$ predstavlja evklidsko razdaljo med napovedno distribucijo verjetnosti razredov i-tega notranjega modela in napovedno distribucijo verjetnosti razredov skupinskega modela.

Lokalno prečno preverjanje (LCV)

Metoda poda oceno zanesljivosti na podlagi meritev napake napovedi pri spremembi okolice. Najprej dobi okolico tako, da poišče K najbližjih sosedov (definicija razdalje med primeroma je lahko poljubna). Nato začne izvajati preverjanje izpusti enega, ki pri vsakem preverjanju ustvari model naučen na vseh najbližjih sosedih, razen tistega, ki je bil v tem testiranju izpuščen. Slednjega namreč nato z modelom napove in izračuna absolutno napako napovedi. Na podlagi napake in razdalje poda oceno zanesljivosti. Celoten postopek je opisan v algoritmu 1.

Algoritem 1 Metoda LCV

procedure LCV(M, x, u)
  ▷ M predstavlja model
  ▷ x je primer za katerega želimo oceniti zanesljivost napovedi
  ▷ u so primeri za katere poznamo vrednosti ciljne spremenljivke
  $knns \leftarrow$ get_knns(u, x, K)
  for each k in $range(len(knns))$ **do**
    $c \leftarrow train(M, knns \backslash knns_k)$
    $E_k \leftarrow |real\_value(knns_k) - c(knns_k)|$
  $ocena\_zanesljivosti \leftarrow \frac{\sum_{k=1}^{K} dist(knns_k,x)*E_k}{\sum_{k=1}^{K} dist(knns_k,x)}$

Na podatkovnih množicah z diskretno ciljno spremenljivko je bila napaka definirana kot

$$P_k = \langle\, I(class\_value(k_i) = val) \,|\, val = class\_value \,\rangle$$

$$E_k = hellinger\_dist(P(C_i), P_k),$$

kjer I predstavlja indikator, *hellinger_dist* predstavlja funkcijo za izračun Hellingerjeve razdalje, $P(C_i)$ pa predstavlja napovedno distribucijo verjetnosti razredov modela naučenega na sosedih brez i-tega soseda. Ocena zanesljivosti je v tem primeru definirana kot

$$ocena\_zanesljivosti = 1 - LCV(M, x, u).$$

V naših testiranjih smo uporabili $K = max(5, st\_primerov/20)$.

Ocenjevanje gostote (DENS)

Zanesljivost napovedi lahko podamo glede na gostoto primerov okoli primera, ki ga napovedujemo. Smiselno je, da je zanesljivost večja, če je okolica primera gosteje poseljena in manjša v primeru, ko je redkeje. Metoda meri gostoto z uporabo Parzenovega okna z Gaussovim jedrom. Gostota okoli primera je definirana kot:

$$gostota(x) = \frac{\sum_{i=1}^{N} K(D(x, x_i))}{N},$$

kjer K predstavlja jedrno funkcijo, D predstavlja funkcijo za izračun razdalje med dvema vektorjema, x je primer, za katerega želimo podati oceno zanesljivosti, N pa število vseh primerov. Ocena zanesljivosti je izražena kot inverz gostote zato, da imamo pozitivno korelacijo z napako napovedi:

$$ocena\_zanesljivosti = max(gostota(x_i)) - gostota(x)$$

V naših poizkusih smo uporabljali Gaussovo jedro in Evklidsko razdaljo.

Lokalno modeliranje napovedne napake (CNK)

Lokalno modeliranje napovedne napake poda oceno zanesljivosti napovedi glede na prave vrednosti K najbližjih sosedov primera, ki ga ocenjujemo. Ocena zanesljivosti je definirana kot

$$CNK(x) = \frac{\sum_{i=1}^{K} class(knns_i(x))}{K} - napoved(x),$$

torej izračuna povprečno vrednost ciljne spremenljivke K najbližjih sosedov ter nato odšteje napovedano vrednost.

Na podatkovnih množicah z diskretno ciljno spremenljivko je ocena zanesljivosti definirana kot 1 - povprečna Hellingerjeva razdalja med napovedno distribucijo verjetnosti razredov primera in napovedno distribucijo verjetnosti razredov najbližjih sosedov.

Kombinacija BAGV + CNK (BVCK)

Kombinacija metod BAGV in CNK se je v preteklih poizkusih izkazala za dobro [2]. Definirana je kot povprečje obeh ocen zanesljivosti:

$$BVCK = \frac{BAGV + CNK}{2}$$

Notranje prečno preverjanje

Metoda s pomočjo prečnega preverjanja na učni množici ugotovi, katera izmed metod ocenjevanja zanesljivosti se najboljše obnese na dani podatkovni množici in izbere najboljšo. Pri množicah z zvezno ciljno spremenljivko se uspešnost meri glede na Pearsonovo korelacijo med napako in oceno zanesljivosti.

Metode nismo uporabili pri testiranjih, saj je zaradi notranjega prečnega preverjanja izrazito počasnejša in bi posledično testiranje vzelo preveč časa.

Referenčna ocena (ORef)

Pevec, Štrumbelj in Kononenko predstavijo novo metodo imenovano referenčna ocena, ki deluje na klasifikacijskih problemih [12]. Za nek primer x, kjer je vrednost ciljne spremenljivke y vemo, da je njegova prava pogojna verjetnost $P(Y = y|x)$. Če bi imeli optimalen model, bi verjetnost, ki nam jo model vrne bila enaka pravi pogojni verjetnosti oziroma bolj natančno, verjetnost bi bila 1 pri razredu, ki je pravilen in 0 pri ostalih razredih. Napako verjetnosti lahko predstavimo z Laplacovo napako

$$f(x) = |prava\_verjetnost - napovedana\_verjetnost(x)|,$$

kjer napovedano verjetnost (v nadaljevanju označena s $\hat{P}(x)$) poda naš klasifikacijski model za izglasovan razred. Ker vemo, da je prava verjetnost enaka 1, je torej napaka $|1 - napovedana\_verjetnost(x)|$, ko določimo pravilno vrednost ciljne spremenljivke in $napovedana\_verjetnost(x)$, ko zgrešimo napoved. Pričakovana vrednost funkcije f je

$$E[f(x)] = P(Y = y|x)(1 - \hat{P}(x)) + (1 - P(Y = y|x))\hat{P}(x).$$

Prava vrednost pogojne verjetnosti nam ni znana, zato upamo, da je $\hat{P}(x)$ dovolj dober približek temu, prav tako pa ne vemo, kateri razred je pravilen in se tudi tu odločimo zaupati našemu modelu in sicer predpostavimo, da je pravilen razred tisti, ki je bil napovedan. Ocena zanesljivosti je torej

$$ocena\_zanesljivosti = \hat{P}(x)(1 - \hat{P}(x)) + (1 - \hat{P}(x))\hat{P}(x).$$

Če se katera izmed metod odreže slabše kot referenčna ocena, jo lahko smatramo kot neuporabno.

Poglavje 3

Ocenjevanje zanesljivosti skupinskih modelov

V nadaljevanju bomo predstavili metode, ki smo jih razvili. Medtem ko so nekatere metode primernejše za klasifikacijske probleme in druge za regresijske, smo želeli ideje metod prenesti v njim manj naravno okolje. Metod Variance notranjih modelov in Delež izglasovanega razreda nismo preslikali, saj smo mnenja, da nismo našli dovolj dobre preslikave. Najprej bomo predstavili tiste, ki so bile razvite za delovanje na podatkih, kjer je ciljna spremenljivka zvezna. Nato bomo predstavili metode razvite za delovanje pri ciljnih spremenljivkah z diskretno vrednostjo.

3.1 Regresijske metode

3.1.1 Varianca notranjih modelov (VAR)

Napoved lahko ocenimo kot zanesljivo, če so notranji modeli imeli zelo podobne napovedi. V kolikor se napovedi precej razlikujejo, lahko neenotnost interpretiramo kot negotovost napovedi. Metodo, ki kot oceno zanesljivosti vrne varianco notranjih modelov, je že preizkusil Močnik [9], ki jo je uporabil na naključnih gozdovih.

3.1.2 K najbližjih stanj odstopanj (KNSO)

Metoda deluje tako, da s skupinskim modelom najprej napovemo testni primer. S tem dobimo napovedi notranjih modelov in skupno napoved skupinskega modela (v nadaljevanju označena s S). Odstopanje notranjega modela smo definirali s formulo $|S - C_iNapoved|$, kjer $C_iNapoved$ predstavlja napoved i-tega notranjega modela. Stanje odstopanj predstavlja množico odstopanj notranjih modelov, kot kaže Slika 3.1. Ocena zanesljivosti je podana na podlagi najbližjih sosedov primera, kjer se sosednost določi z razdaljo med dvema stanjema odstopanj. Razdaljo med primeroma smo definirali z enačbo (3.1) v kateri sta a in b vektorja med katerima računamo razdaljo, n pa predstavlja velikost vektorja.

$$dist(a, b) = \sum_{i=1}^{n} |a_i - b_i| \tag{3.1}$$

Za vsak primer vemo, kakšna je bila napaka napovedi skupinskega modela pri napovedi njegove ciljne spremenljivke, saj smo ga pridobili iz učne množice. Metoda oceno zanesljivosti izračuna tako, da vzame povprečno napako k najbližjih sosedov. Zaradi predpostavke, da imajo bližji sosedi bolj podobno napako kot tisti, ki so bolj oddaljeni, se pri računanju ocene upošteva tudi razdaljo soseda, kar je razvidno iz enačbe (3.2). Ta je bila uporabljena tudi v naših poizkusih. Razdalja sosedom poveča napako glede na njihovo oddaljenost.

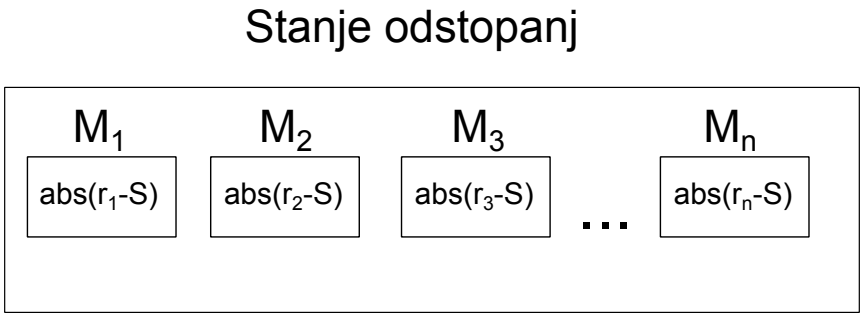
$$ocena\_zanesljivosti = \frac{\sum_{i=1}^{k} napaka_i + distance_i}{k} \tag{3.2}$$

Preizkusili smo različne definicije razdalj med katerimi je bila tudi evklidska, vendar smo dobili slabše rezultate kot z zgoraj opisano definicijo razdalje. Celoten postopek je opisan v Algoritmu 2.

Metoda KNSO napoved notranjih modelov primerja z napovedjo skupinskega modela in na podlagi tega izračuna odstopanje notranjega modela, ki je v zveznem prostoru absolutna vrednost razlik napovedi. V diskretnem prostoru je odstopanje še lažje definirati. Določimo ga tako, da preverimo enakost napovedi notranjega in skupinskega modela. Stanje odstopanj je binaren zapis, na podlagi katerega bomo iskali najbližje sosede glede na Hammingovo razdaljo. To nas lahko zavede, da je preslikava enaka kot metoda Knn notranjih modelov (KnnNM), pri kateri smo na enak način merili razdaljo sosedov, vendar se metodi razlikujeta v tem,

Algoritem 2 Metoda KNSO

procedure KNSO(c, t, u)
    ▷ c predstavlja naš skupinski model
    ▷ t je primer za katerega želimo oceniti zanesljivost napovedi
    ▷ u je učna množica na kateri se je c naučil
    $S \leftarrow \mathrm{c}(t)$
    $stanje\_odstopanj \leftarrow |S - C_i Napoved|$ for C_i in notranji_modeli
    $knns \leftarrow$ get_knns(stanje_odstopanj, c, u, k=5)
    $cnt \leftarrow 0$
    $vsota \leftarrow 0$
    for each *neighbour* in $knns$ **do**
        $vsota \leftarrow vsota + (error(neighbour) + distance(neighbour))$
        $cnt \leftarrow cnt + 1$
    $ocena\_zanesljivosti \leftarrow$ vsota/cnt
procedure GET_KNNS(r, c, u, $k = 5$)
    $stanja \leftarrow$ empty array
    for each t in u **do**
        $S \leftarrow \mathrm{c}(t)$
        $cur\_stanje \leftarrow |S - C_i Napoved|$ for C_i in notranji_modeli
        $abs\_napaka \leftarrow |class(t) - S|$
        add $(cur\_stanje, abs\_napaka, dist(r, cur\_stanje))$ to $stanja$
    $knns \leftarrow$ take first k elements from sort(stanja by dist)



Slika 3.1: Prikaz stanja odstopanj, kjer M_i predstavlja i-ti notranji model, r_i predstavlja napoved i-tega notranjega modela in S predstavlja napoved skupinskega modela.

da medtem ko KnnNM primerja enakost istoležnih napovedi notranjih modelov dveh primerov, KNSO primerja enakost napovedi notranjih modelov z napovedjo skupinskega modela.

3.1.3 Stacking notranjih modelov (SNM)

Ideja metode izhaja iz modela imenovanega stacking. Stacking deluje tako, da imamo nekaj primarnih modelov, s katerimi ocenimo dani primer, ter njihove napovedi zabeležimo v tabelo velikosti $M * (N + 1)$, kjer je M število učnih primerov in N število primarnih modelov. Vrstica v tabeli poleg napovedi modelov vsebuje tudi pravo vrednost razredne spremenljivke, zaradi česar jo lahko uporabimo kot učno množico. Na tej tabeli naučimo sekundarni model, recimo mu C, s katerim podamo končno napoved. Za napoved testnega primera je potrebno primer najprej pretvoriti v obliko, kot je predstavljena v novi učni množici, torej ga je potrebno najprej oceniti s primarnimi modeli. Ko to storimo, lahko C poda svojo napoved.

Kot omenjeno je ideja te metode zelo podobna. Za primarne modele smo vzeli notranje modele in iz njihovih napovedi sestavili tabelo napovedi. Primeri, iz katerih smo sestavili tabelo, so del učne množice, na kateri so bili notranji modeli naučeni, kar ni optimalno, saj so se lahko modeli prilagodili učni množici.

Bolje bi bilo napovedovati primere, na katerih niso bili naučeni, vendar metoda predpostavi, da dobimo le en testni primer in ker imamo že napovedano vrednost skupinskega modela pomeni, da je že naučen na učni množici. Za razliko od modela stacking, tej tabeli ne dodamo prave vrednosti razredne spremenljivke, ampak napako skupinske napovedi, ki je za zvezno spremenljivko definirana kot absolutna razlika med napovedjo skupinskega modela in prave vrednosti. Pravo vrednost poznamo, saj je primer del učne množice. Podobno kot zgoraj opisana metoda KNSO, tudi ta poizkusi uganiti napako, vendar ne na podlagi sosedov primera temveč z napovedjo sekundarnega modela. Celoten postopek je opisan v Algoritmu 3.

Algoritem 3 Metoda Stacking notranjih modelov

procedure SNM(c, t, u)
    ▷ c predstavlja naš skupinski model
    ▷ t je primer za katerega želimo oceniti zanesljivost napovedi
    ▷ u je učna množica na kateri se je c naučil
    $nova\_ucna \leftarrow$ zgradi_tabelo_rezultatov(u, c)
    $sekundarni\_model \leftarrow$ sekundarni_model.learn($nova\_ucna$)
    $ocena\_zanesljivosti \leftarrow$ sekundarni_model(t)
procedure ZGRADI_TABELO_REZULTATOV(u, c)
    $tabela \leftarrow$ empty array
    for each t in u **do**
        $S \leftarrow$ c(t)
        $NS \leftarrow$ cur_c(t) for cur_c in notranji_modeli
        $abs\_napaka \leftarrow |S - class(t)|$
        add $[NS + abs\_napaka]$ to $tabela$
    $tabela$

Naravna preslikava metode za reševanje klasifikacijskih problemov bi v tabeli napovedi imela napako predstavljeno kot binarno vrednost, ki bi nam povedala ali je skupinski model v tem primeru pravilno ali napačno napovedal. S tem sicer ni nič narobe, ampak kot ocena zanesljivosti ni primerna, saj bi za ovrednotenje uspešnosti metode z Wilcoxon-Mann-Whitneyevo statistiko za smiselne rezultate

potrebovali zvezne vrednosti. Večina modelov lahko poleg napovedi poda še njeno verjetnost. Skupinski modeli, ki so bili uporabljeni, so lahko podali verjetnost s pomočjo katere lahko definiramo oceno zanesljivosti kot $1 - P(Zadel)$. Preslikana različica metode je, ob predpostavki, da sta napaka in verjetnost napovedi korelirani, dovolj dober približek originalni. V naših poizkusih smo za sekundarni model uporabili naključne gozdove.

3.2 Klasifikacijske metode

3.2.1 Delež izglasovanega razreda (Ratio)

Prva ideja, ki smo jo dobili za oceno zanesljivosti napovedi klasifikacijskega modela, je bila uporaba deleža notranjih klasifikatorjev, ki so napovedali napovedani razred. V kolikor je bilo notranjih modelov 50 in je izmed teh 32 glasovalo za razred, ki je bil izglasovan, potem je delež enak 32/50. Oceno zanesljivosti predstavlja omenjeni delež.

$$ocena\_zanesljivosti = \frac{\#i : C_i = S}{\#i},$$

kjer je C_i napoved i-tega notranjega modela in S skupna napoved skupinskega modela.

3.2.2 Klasifikacijska preciznost (Prec)

Glede na to, da je skupinski model napovedal diskretno vrednost (v nadaljevanju označena z X), je smiselno sklepati, da je X napovedalo nekaj notranjih klasifikatorjev. Ravno tej klasifikatorji so uspeli prepričati skupinski model k takšni napovedi. Obstaja možnost, da kateri izmed notranjih klasifikatorjev, ki je napovedal X, o X-u v resnici ne ve veliko in je napoved kvečjemu rezultat ugibanja. V tem primeru njegova napoved ni zanesljiva. Za vsakega izmed notranjih klasifikatorjev, ki so napovedali X, bi radi vedeli kako zanesljiv je pri napovedi tega razreda. Potrebno je torej ugotoviti delež pravilnih napovedi posameznega notranjega modela pri napovedi danega razreda. Gre torej za ugotavljanje preciznosti notranjih modelov pri danem razredu,

$$preciznost = TP/(TP + FP),$$

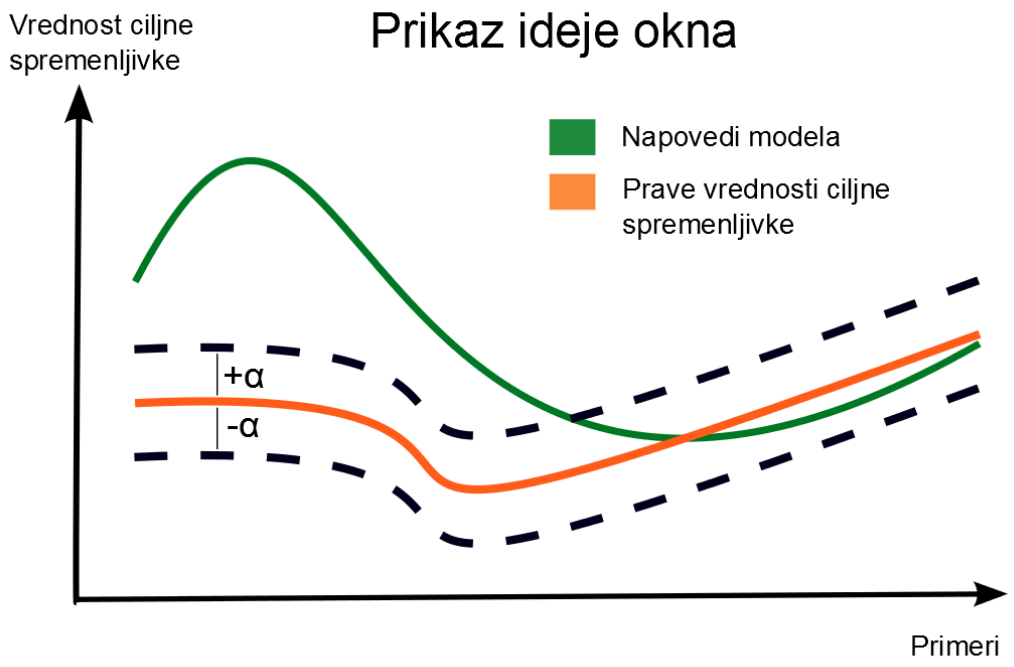
kjer TP predstavlja število pravilnih napoved in FP število napačnih napovedi, ko je model napovedal dani razred.

Metoda, kot je bila uporabljena v poizkusih, je opisana v Algoritmu 4.

Algoritem 4 Metoda Klasifikacijska preciznost

```
1:  procedure PREC(c, t, u)
2:      ▷ c predstavlja naš skupinski model
3:      ▷ t je primer za katerega želimo oceniti zanesljivost napovedi
4:      ▷ u je učna množica na kateri se je c naučil
5:      skupna_napoved ← c(t)
6:      cnt ← 0
7:      vsota ← 0
8:      for each m in notranji_modeli do
9:          if m(t) = skupna_napoved then
10:             preciznost ← preciznost(m, u, skupna_napoved)
11:             cnt ← cnt + 1
12:             vsota ← vsota + preciznost
13:     ocena_zanesljivosti ← vsota/cnt
14: procedure PRECIZNOST(m, u, x)
15:     tp, fp ← 0,0
16:     for each t in u do
17:         napoved ← m(t)
18:         if napoved = x then
19:             if class(t) ≠ napoved then
20:                 fp ← fp + 1
21:             else
22:                 tp ← tp + 1
23:     preciznost ← tp/(tp + fp)
```

Zgoraj opisani način ocenjevanja preciznosti ne upošteva razlike med klasifikatorjem, ki je pravilno napovedal en poizkus od dveh, od tistega, ki je petkrat pravilno napovedal od skupno desetih napovedi. Za slednjega namreč lahko z



Slika 3.2: Prikaz ideje okna, kjer črtkasti krivulji predstavljata okno

večjo zanesljivostjo trdimo, da jih polovico zadane. Zato smo poizkusili namesto povprečne preciznosti notranjih klasifikatorjev seštevati pravilne pozitivne primere (TP) in napačne negativne (FP), a se je metoda slabše obnesla. Poizkusili smo tudi z m-oceno z m vrednostmi 2, 5 in 10, vendar nismo dosegli boljših rezultatov.

Pri preslikavi metode za reševanje regresijskih problemov je potrebno biti pozoren pri računanju preciznosti modelov. Glede na to, da je ciljna spremenljivka zvezna je zelo majhna verjetnost, da bi kateri izmed modelov zadel pravo vrednost in bi v tem scenariju večina modelov imela preciznost 0. Problem lahko rešimo z idejo okna okoli napovedi, kot je prikazano na Sliki 3.2. Na primeru iz slike je model napovedal pravo vrednost na desni strani slike saj je znotraj okna. Pri oknu moramo definirati kakšno je dovoljeno odstopanje. V naših testih smo izbrali 5 odstotkov razpona med najmanjšo in največjo napovedjo v vsako smer. Preciznost notranjega modela dobimo tako, da najprej dobimo absolutne napake napovedi, ko je napoved modela bila znotraj okna in nato izračunamo povprečje teh napak. Ko

imamo ta podatek za vse notranje modele, vzamemo povprečje in to predstavlja oceno zanesljivosti.

$$okno = (skupna\_napoved - delta, skupna\_napoved + delta)$$

3.2.3 KNN notranjih modelov (KnnNM)

Metoda KnnNM temelji na ideji iskanja najbližjih sosedov. Najbližje sosede smo iskali v učni množici. Pri določanju sosedov je najprej potrebno definirati vektorje med katerimi bomo merili razdalje in način izračuna razdalje. Definirajmo vektorja:

$$a = [X, Y, X, Z, X, Z]$$

$$b = [X, X, X, Z, X, Y]$$

Vrednosti vektorjev predstavljajo napovedi notranjih modelov (v tem primeru je bilo notranjih modelov 6). Razdaljo med dvema vektorjema smo merili z Hammingovo razdaljo, ki prešteje, na koliko istoležnih elementih se vrednosti razlikujeta (razdalja med vektorjema a in b je torej 2). Za vsakega soseda imamo tudi podatek, ali je bila pri njem napoved skupinskega klasifikatorja pravilna ali napačna. Ocena zanesljivosti je bila definirana glede na razdaljo sosedov in pravilnost napovedi skupinskega klasifikatorja pri njihovem primeru. Pomembnost soseda je morala biti odvisna od njegove razdalje, za to smo poskrbeli tako, da je sosed prispeval

$$\frac{1}{dist(primer, sosed)}$$

h končnemu rezultatu. Potrebno je bilo tudi upoštevati ali je pri tem sosedu skupinski klasifikator podal pravilno napoved. To smo rešili tako, da v primeru pravilnega rezultata prispevek soseda prištejemo k oceni zanesljivosti, v primeru napačnega pa odštejemo. Celoten postopek je opisan v Algoritmu 5.

Algoritem 5 Metoda KNN notranjih modelov

```
1: procedure KnnNM(c, t, u)
2:     ▷ c predstavlja naš skupinski model
3:     ▷ t je primer za katerega želimo oceniti zanesljivost napovedi
4:     ▷ u je učna množica na kateri se je c naučil
5:     skupna_napoved, napovedi_notranjih_modelov ← c(t)
6:     knns ← get_knns(napovedi_notranjih_modelov, u, c)
7:     eps ← 1E-8
8:     ocena_zanesljivosti ← 0.0
9:     for each (dist, zadel) in knns do
10:        if zadel then
11:            ocena_zanesljivosti ← ocena_zanesljivosti + 1/(dist + eps)
12:        else
13:            ocena_zanesljivosti ← ocena_zanesljivosti - 1/(dist + eps)
14:    ocena_zanesljivosti
15: procedure get_knns(state, u, c)
16:    knns ← empty array
17:    for each primer in u do
18:        cur_stanje ← cl(primer) for cl in notranji_modeli(c)
19:        skupna_napoved ← c(primer)
20:        zadel ← skupna_napoved = class(primer)
21:        add (hamming(stanje, cur_stanje), zadel) to knns
22:    knns
```

S podobnim problemom kot pri preslikavi metode Klasifikacijska preciznost, se srečamo tudi pri preslikavi KnnNM in sicer pri merjenju razdalje dveh vektorjev. Metoda meri Hammingovo razdaljo med napovedmi notranjih modelov primera, ki ga ocenjujemo in napovedmi ostalih primerov. Pri zveznih vrednostih bosta vrednosti zelo redko enaki in bi večina razdalj bilo enakih maksimalni razdalji, ki je enaka številu notranjih modelov. Tako kot pri metodi Klasifikacijska preciznost lahko tudi tu uporabimo idejo okna okoli napovedi. Velikost okna smo vzeli enako kot pri pretvorbi Klasifikacijske preciznosti, torej v vsako smer 5 odstotkov

razpona med najmanjšo in največjo napovedjo. Pri računanju Hammingove razdalje označimo istoležna elementa enaka v kolikor je napoved prvega znotraj okna napovedi drugega.

Poglavje 4

Postopek testiranja metod

Vse metode smo testirali z 10-kratnim preverjanjem. Učno množico posameznega preverjanja je sestavljalo 75 odstotkov naključno izbranih primerov testno pa preostalih 25 odstotkov. Metode, ki delujejo na regresijskih primerih smo primerjali med seboj glede na Pearsonovo korelacijo med napako napovedi in oceno zanesljivosti. Pearsonovo korelacijo bi lahko za vsako testno množico izračunali posamično in nato uspešnost metode predstavili z njenim povprečjem, vendar smo jo raje izračunali na koncu, ko smo končali 10 kratno preverjanje in imeli za testne primere tudi podano oceno zanesljivosti. Pri klasifikacijskih primerih smo uspešnost metod ocenili z Wilcoxon-Mann-Whitneyevo statistiko. Tako kot pri Pearsonovi korelaciji, smo tudi Wilcoxon-Mann-Whitneyevo statistiko izračunali le enkrat in sicer takrat, ko smo imeli podane ocene zanesljivosti za vse testne primere. Pri testiranju podatkovnih množic z diskretno ciljno spremenljivko smo poskrbeli, da imajo vse učne in testne množice enako distribucijo vrednosti razredne spremenljivke.

Pri implementaciji smo uporabili programski paket Orange [7]. Pri skupinskem modelu bagging smo za model izbrali SimpleTreeLearner, ki predstavlja regresijsko oziroma odločitveno drevo. Skupinske modele smo testirali s 50 in 100 notranjimi modeli. Omenjene metode iz poglavja 2, ki so neodvisne od modela se nahajajo v dodatku za modul Orange imenovanim Orange-reliability [15], katerega smo tudi uporabili pri testiranju.

Tabela 4.3 Prikaz podrobnosti klasifikacijskih podatkovnih množic

Ime	Primerov	Atributov
anneal	898	39
audiology	226	70
balance-s	625	5
breast	286	10
bcwd	683	10
bupa	345	7
crx	690	16
glass	214	10
horse-colic	368	21
ionosphere	351	33
iris	150	5

Ime	Primerov	Atributov
lymph	148	19
monks-1	556	7
monks-2	601	7
monks-3	554	7
post-del	90	9
primary	339	18
shuttle	253	7
tic_tac_toe	958	10
voting	435	17
wine	178	14
zoo	101	17

4.1 Podatkovne množice

Metode smo testirali na podatkovnih množicah z zvezno ciljno spremenljivko, kot tudi z diskretno. V tem poglavju bomo predstavili podatkovne množice na katerih smo metode testirali. Najprej bomo predstavili podatkovne množice z diskretno ciljno spremenljivko, nato bo sledila še predstavitev množic z zvezno ciljno spremenljivko.

Klasifikacijske podatkovne množice

Pri testiranju metod, ki lahko ocenijo zanesljivost napovedi klasifikacijskega primera smo uporabili podatkovne množice, ki se nahajajo znotraj programskega paketa Orange. Metode smo testirali na 22 podatkovnih množicah, katerih število primerov in atributov opisuje Tabela 4.3.

Tabela 4.6 Prikaz podrobnosti regresijskih podatkovnih množic

Ime	Primerov	Atributov
ACE	114	192
AChE	111	143
AMPH1	130	85
BZR	163	240
EDC	123	178

Ime	Primerov	Atributov
HIVPR	113	156
HIVRT	101	123
h-PTP	135	170
DHFR	397	212
COX2	322	252

Regresijske podatkovne množice

Metode, ki znajo oceniti regresijsko napoved smo testirali na delu podatkovnih množicah, ki so bile uporabljene v raziskavi [16]. Pri naših testih so množice vsebovale le atribute, ki opisujejo krožni odtis molekule, saj je bilo v omenjenem članku pokazano, da ni bistvenih razlik v rezultatih med množicami, ki so vsebovale samo informacije o krožnem odtisu, množicami, ki so vsebovale le atribute ustvarjene z orodjem RDKit in tistimi, ki so vsebovale atribute obeh vrst. Podatkovne množice, ki jih je izdelalo podjetje Astra Zeneca, vsebujejo podatke o kvantitativnih razmerjih med strukturo in delovanjem molekule. Uporabili smo 10 podatkovnih množic, ki so podrobneje opisane v Tabelah 4.6 in 4.7.

4.2 Mere statistike

Pearsonova korelacija

Pearsonov korelacijskih koeficient dveh spremenljivk nam pove, kako močno sta spremenljivki povezani. Korelacija je pozitivna takrat, ko se ob povečanju ene spremenljivke poveča tudi druga, negativna pa ko sta si spremenljivki obratno sorazmerni. Korelacija vedno zavzema vrednost na intervalu od -1 do 1. Pearsonov korelacijski koeficient je definiran kot razmerje kovariance (4.2) spremenljivk in produkta njunih standardnih odklonov, kar je razvidno iz enačbe (4.1).

$$r(a,b) = \frac{kovarianca(a,b)}{\sigma_a \sigma_b} \tag{4.1}$$

Tabela 4.7 Prikaz podrobnosti regresijskih podatkovnih množic

Ime	Opis
ACE	inhibition of angiotensin-converting enzyme
AChE	inhibition of acetylcholinesterase
AMPH1	binding affinity to the human amphiphysin-1 SH3 domain
BZR	binding affinity to the benzodiazepine receptor
EDC	relative binding affinity to the estrogen receptor
HIVPR	inhibition of human immunodeficiency virus protease
HIVRT	inhibition of HIV-1 reverse transcriptase
h-PTP	inhibition of human protein tyrosine phosphatase 1B
DHFR	inhibition of dihydrofolate reductase
COX2	inhibition of cyclo-oxygenase-2

$$kovarianca(a,b) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} \tag{4.2}$$

Pomemben je še izračun P-vrednosti. Ta nam pove, kakšna je verjetnost, da pridemo do takšnega Pearsonovega korelacijskega koeficienta v primeru, ko spremenljivki nista povezani. S pomočjo P-vrednosti, lahko ocenimo v katerih primerih so ocene metod statistično značilne.

Wilcoxon-Mann-Whitneyeva U statistika

Omenili smo že, da v klasifikacijskih primerih napaka zavzema binarno vrednost, ki nam pove ali je klasifikator zadel ali zgrešil. Torej imamo dve skupini primerov za katere bomo podali oceno zanesljivosti. Želeli bi izvedeti ali sta skupini enaki ali različni. Wilcoxon-Mann-Whitneyevo statistika je neparametričen test, ki testira ničelno hipotezo, da ni razlike med populacijama. Najprej uredi binarne vrednosti naraščajoče glede na njihove zvezne vrednosti (v našem primeru so to ocene zanesljivosti). Pri urejenem seznamu za vsak element prešteje, koliko nasprotnih vrednosti ima na svoji levi. Naj $U(0)$ predstavlja vsoto vrednosti pri populaciji z oznako '0' in $U(1)$ vsoto vrednosti pri populaciji z oznako '1'. U vrednost je de-

finirana kot $min(U(0), U(1))$. Na podlagi U vrednosti in velikosti populacij nato izračuna kako statistično značilen je rezultat. Formula za izračun U vrednosti je

$$U1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

kjer n_1 in n_2 predstavljata velikosti populacij, R_1 in R_2 predstavljata vsoto rangov primerov posamezne populacije, vmes pa je formula za izračun prvih N zaporednih števil. Pri testiranju smo uporabili implementacijo iz paketa SciPy [8].

4.3 Uporabljeni skupinski modeli

V nadaljevanju bomo opisali skupinska modela, ki sta bila uporabljena pri testiranju. Vsakega izmed skupinskih modelov smo testirali na 50 in 100 notranjih modelih.

Bagging

Ideja bagginga [4] je, da ustvarimo M notranjih modelov. Za vsak notranji model sestavimo množico, na kateri se bo naučil tako, da iz učne množice vzamemo N (število primerov v učni množici) naključnih primerov s ponavljanjem. S tem dobimo M različnih modelov in na podlagi njihovih napovedi skupinski model poda končno napoved. Glasovanje modelov ni uteženo, kar pomeni, da so glasovi enakovredni. Pri diskretnih primerih je skupna napoved razred, ki je prejel največ glasov, za zvezne primere pa se izračuna povprečna napoved notranjih modelov, kot prikazuje enačba (4.3).

$$C = \frac{\sum_{i=1}^{M} C_i}{M} \tag{4.3}$$

Naključni gozdovi

Naključen gozd [5] sestavlja množica dreves. V kolikor imamo klasifikacijski problem, so to odločitvena drevesa, v regresijskih problemih pa regresijska drevesa.

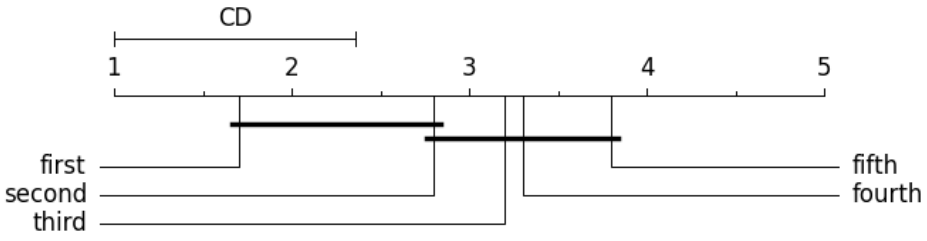
Drevo v naključnem gozdu se razlikuje v tem, da pri gradnji drevesa prihaja do naključnih odločitev. Naj bo N število učnih primerov in M število atributov v podatkovni množici. Pri gradnji posameznega drevesa najprej sestavimo učno množico velikosti N, na kateri se bo učil, tako, da naključno izberemo N primerov iz učne množice s ponavljanjem. Na vsakem vozlišču pri izbiri atributa, glede na katerega se bo drevo vejilo, izberemo m naključnih atributov iz učne množice, med katerimi izberemo najboljšega. Parameter m je za vsako drevo enak in velja, da je manjši od M. V primeru klasifikacije je napoved naključnega gozda najpogosteje napovedan razred, pri regresiji pa povprečje napovedanih vrednosti.

4.4 Primerjava uspešnosti metod

Omenili smo, da bomo uspešnost metod na regresijskih problemih merili s Pearsonovim korelacijskih koeficientom med absolutno napako napovedi in oceno zanesljivosti, na klasifikacijskih pa z Wilcoxon-Mann-Whitneyevo statistiko. V delu želimo metode primerjati z obstoječimi, torej želimo metode razvrstiti glede na uspešnost. To lahko storimo z grafom rangov s kritično razdaljo, ki nam pove tudi ali je bila katera izmed metod statistično značilno boljša od druge.

Graf rangov s kritično razdaljo

Primerjanje metod lahko izvedemo tako, da pri vsaki podatkovni množici razvrstimo metode glede na njihovo uspešnost od najbolj uspešne metode (ta bo zasedla prvo mesto) do najmanj uspešne (zadnje mesto). Enostavna in smiselna zamisel določanja, kako dobro se je posamezna metoda odrezala je, da vzamemo njeno povprečno uvrstitev glede na rezultate pri vseh množicah.



Slika 4.1: Primer grafa rangov s kritično razdaljo

Graf rangov 4.1 [6] nam ravno to razvrstitev prikaže in sicer na levi strani imamo najuspešnejše metode in bolj ko gremo na desno stran, manj so te metode uspešne. Zanimiv podatek je tudi ali so si metode statistično značilno različne med seboj. Nemenyijev test, prikazan v enačbi (4.4), nam pri podanem številu podatkovnih množic (N) in številu metod oziroma algoritmov (K) lahko določi razdaljo za katero se morata algoritma razlikovat, da ju lahko označimo kot statistično značilno različna. Tej razdalji pravimo kritična razdalja. Pri poizkusih smo izbrali parameter stopnje zaupanja (α) enak 0,05.

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}} \tag{4.4}$$

Poglavje 5

Rezultati

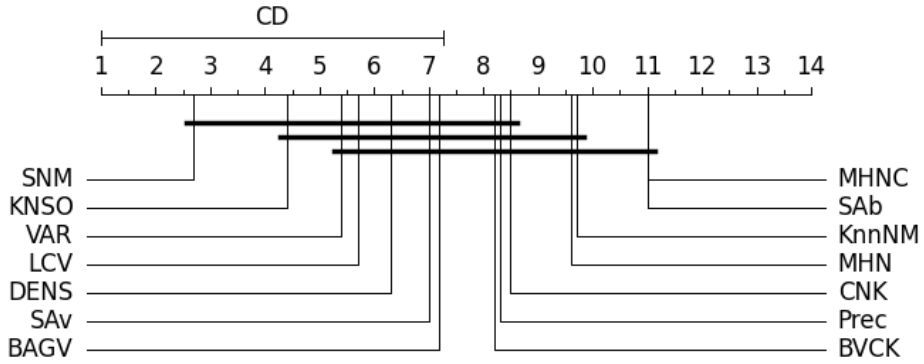
V tem poglavju bomo predstavili rezultate testiranj metod. Pri vsakem skupinskem modelu smo uspešnost metod predstavili z grafom kritične razdalje, ki mu sledi kratka analiza njegovih rezultatov ter podrobnejša predstavitev rezultatov metod pri posamezni podatkovni množici.

5.1 Regresijske podatkovne množice

Na podatkovnih množicah z zvezno ciljno spremenljivko smo testirali metode z dvema skupinskima modeloma in sicer z naključnimi gozdovi in modelom bagging. Tabele rezultatov so prikazane tako, da stolpci predstavljajo podatkovne množice. Posamezna vrstica predstavlja rezultate določene metode na teh podatkovnih množicah. Rezultate, ki so bili statistično značilni, smo prikazali z odebeljeno pisavo. Stolpce tabele smo uredili od leve proti desni glede na število metod, ki so pri podatkovni množici dosegle rezultat, ki je statistično značilen. Skrajno levi stolpec torej predstavlja podatkovno množico, na kateri je največ metod doseglo statistično značilen rezultat, medtem ko skrajno desni stolpec predstavlja množico, na kateri je najmanj metod prišlo do statistično značilnega rezultata. Tudi metode so razvrščene glede na število podatkovnih množic, na katerih so dosegle statistično značilen rezultat. Metoda, katere rezultat je bil pri največ množicah statistično značilen, je v prvi vrstici, metoda, ki je najmanjkrat prišla do statistično značilnega rezultata, pa v zadnji.

5.1.1 Naključni gozdovi

Razvrstitev metod glede na uspešnost pri naključnih gozdovih s 50 drevesi kaže Slika 5.1. Nove metode, ki so bile razvite za podatkovne množice z zvezno ciljno spremenljivko, so se odrezale boljše v primerjavi s starimi, saj so na grafu na levi strani. Metodi Prec in KnnNM, ki sta bili prilagojeni za delo z zveznimi spremenljivkami, se nista preveč dobro izkazali in sta na desni strani, torej med slabšimi. Pomembno je poudariti, da nobena izmed novih metod ni statistično značilno boljša od starih metod. Razlogov za to je lahko več. Morda smo testirali na premajhnem številu podatkovnih množic in bi v primeru večjega števila lahko potrdili razliko med metodami, saj bi kritična razdalja postala krajša.



Slika 5.1: Prikaz uspešnosti metod z grafom kritične razdalje pri naključnih gozdovih s 50 drevesi

Iz Tabele 5.1 je razvidno, da je metoda SNM edina prišla do statistično značilnega rezultata pri vseh množicah, kar pomeni, da je verjetnost, da pridemo do takega rezultata z naključjem, manjša od 5%. Najslabša metoda je dosegla statistično značilen rezultat na treh množicah. Razlog za to je najbrž v tem, da smo rezultate posameznih testiranj združili in izračunali Pearsonov korelacijski koeficient iz precej velikega števila točk (okoli 300 v povprečju). Pri tako velikem vzorcu je mogoče tudi za rezultate, ki niso tako ekstremni, pokazati, da najbrž niso takšni zaradi naključja. Rezultati so zanimivi tudi zaradi tega, ker so koeficienti nizki, saj običajno ko govorimo o korelaciji mislimo na močnejšo korelacijo. Povsod gre torej za šibko korelacijo na teh podatkovnih množicah. Metode smo rangirali glede na

Tabela 5.1 Prikaz Pearsonovih koeficientov metod na posameznih podatkovnih množicah pri skupinskem modelu naključnih gozdov s 50 notranjimi modeli. Podatkovne množice so opisane v Tabeli 4.6

Metoda	ACHE	DHFR	COX2	BZR	ACE	HIVPR	HPTP	EDC	AMPH1	HIVRT
SNM	**0.39**	**0.37**	**0.27**	**0.35**	**0.19**	**0.18**	**0.30**	**0.18**	**0.23**	**0.49**
KNSO	**0.34**	**0.35**	**0.17**	**0.28**	**0.24**	**0.27**	**0.19**	**0.15**	**0.27**	0.04
VAR	**0.39**	**0.24**	**0.13**	**0.37**	**0.18**	**0.24**	**0.19**	0.04	**0.24**	−0.08
SAv	**0.43**	**0.24**	**0.08**	**0.16**	0.07	**0.15**	**0.17**	−0.09	**0.13**	**−0.20**
DENS	**0.28**	**0.22**	**0.08**	0.08	−0.13	**0.33**	**0.20**	**0.17**	−0.05	**0.27**
LCV	**0.23**	**0.29**	**0.23**	**0.47**	**0.16**	0.12	0.11	**0.18**	0.11	0.08
BAGV	**0.35**	**0.24**	**0.09**	**0.38**	0.11	0.09	**0.18**	0.01	**0.15**	−0.09
Prec	−0.10	0.01	0.05	**0.19**	**0.32**	−0.07	−0.03	**0.22**	**0.21**	**0.19**
MHN	**0.16**	**0.16**	0.03	−0.10	0.07	**0.27**	0.04	**0.15**	−0.01	−0.10
CNK	**0.26**	**0.26**	**0.09**	0.04	**0.25**	0.11	0.08	0.12	−0.11	0.06
MHNC	**0.15**	**0.17**	0.02	−0.06	0.08	**0.24**	0.03	**0.14**	−0.00	−0.05
BVCK	**0.27**	**0.26**	**0.09**	0.04	**0.25**	0.11	0.08	0.12	−0.11	0.05
SAb	**0.32**	**0.10**	0.01	**0.23**	0.03	−0.08	0.05	0.05	0.04	−0.06
KnnNM	0.05	0.02	**−0.10**	−0.07	0.12	0.08	**−0.16**	0.09	0.06	**−0.15**

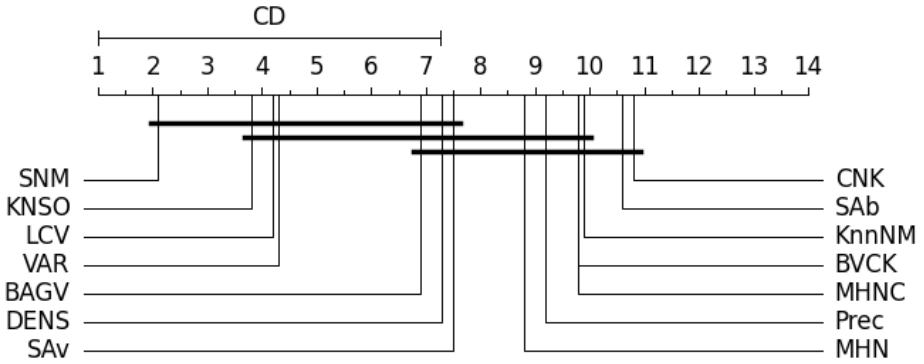
povprečen rang, ki so ga dosegle na podatkovnih množicah. Rang na posamezni podatkovni množici je bil odvisen od njenega Pearsonovega koeficienta (metoda z najvišjim koeficientom ima rang 1, itd.), torej se rezultati tabele ujemajo s sliko uspešnosti metod, saj imajo KNSO, SNM in VAR višje koeficiente v primerjavi z ostalimi metodami. Na dnu tabele sta metodi SAb in KnnNM, saj so njuni koeficienti precej bližje ničli.

Pri 100 notranjih modelih pričakujemo, da se bodo metode, ki upoštevajo notranje modele (SNM, KNSO in VAR), izboljšale, saj imajo več informacij. Zaporedje metod na levi strani grafa se ni bistveno spremenilo, so pa nastale večje razlike med povprečnimi rangi metod kar je razvidno is Slike 5.2. Metoda SNM je izboljšala svojo povprečno uvrstitev za polovico ranga, prav tako so rezultate izboljšale KNSO, LCV in VAR, ki so precej strnjene skupaj okoli četrtega mesta. Metoda SNM je bila statistično značilno boljša od vseh metod na desni strani grafa. Presenetljivo slabo se je odrezala metoda CNK, ki je na skrajni desni strani grafa, in posledično tudi BVCK, saj uporablja njen rezultat. Tudi pri 100 notra-

Tabela 5.2 Prikaz Pearsonovih koeficientov metod na posameznih podatkovnih množicah pri skupinskem modelu naključnih gozdov s 100 notranjimi modeli.

Metoda	DHFR	ACHE	HPTP	COX2	ACE	BZR	AMPH1	HIVPR	EDC	HIVRT
SNM	**0.40**	**0.39**	**0.39**	**0.23**	**0.18**	**0.34**	**0.35**	**0.19**	**0.20**	**0.44**
LCV	**0.36**	**0.37**	**0.20**	**0.18**	**0.16**	**0.37**	**0.18**	**0.15**	**0.18**	0.13
KNSO	**0.38**	**0.39**	**0.26**	**0.09**	**0.18**	**0.31**	**0.38**	**0.26**	**0.13**	0.10
VAR	**0.29**	**0.44**	**0.29**	**0.12**	**0.16**	**0.40**	**0.39**	**0.23**	0.08	0.03
BAGV	**0.25**	**0.32**	**0.17**	**0.08**	**0.17**	**0.36**	**0.21**	0.13	0.05	0.09
DENS	**0.25**	**0.22**	**0.18**	0.06	−0.07	**0.11**	−0.11	**0.30**	0.07	**0.28**
SAb	**0.10**	**0.20**	0.08	**0.09**	**0.15**	**0.18**	0.02	0.02	0.05	0.01
SAv	**0.32**	**0.37**	**0.30**	0.03	0.08	**0.22**	**0.24**	0.08	−0.05	-0.10
MHN	**0.16**	**0.15**	0.02	**0.13**	−0.11	−0.05	−0.04	**0.15**	0.06	−0.11
Prec	0.06	−0.02	−0.00	0.02	**0.28**	**0.25**	**0.32**	−0.03	0.12	0.09
CNK	**0.15**	0.03	**0.14**	0.07	**0.13**	0.06	0.05	0.08	0.04	0.10
MHNC	**0.17**	**0.15**	0.02	**0.13**	−0.10	−0.03	−0.05	0.12	0.06	−0.05
BVCK	**0.15**	0.03	**0.14**	0.07	**0.13**	0.06	0.05	0.08	0.04	0.10
KnnNM	0.03	0.13	**−0.21**	**−0.11**	−0.00	0.04	−0.10	−0.02	**0.14**	−0.08

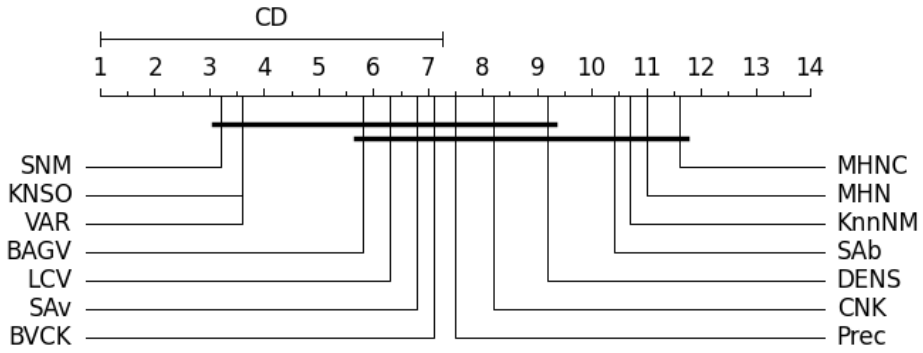
njih modelih je le metoda SNM prišla do statistično značilnega rezultata na vseh množicah (Tabela 5.2). Kar tretjina metod je prišla do statistično značilnega rezultata le na treh množicah. Vrstni red podatkovnih množic se je malo spremenil, sprememba je vidna predvsem na množici HPTP, kjer so bile tri dodatne metode uspešne.



Slika 5.2: Prikaz uspešnosti metod z grafom kritične razdalje pri naključnih gozdovih s 100 drevesi

5.1.2 Bagging

Uspešnost metod pri skupinskem modelu bagging s 50 notranjimi modeli, kaže Slika 5.3. Kot pri grafu rangov kritične razdalje skupinskega modela naključnih gozdov (Slika 5.1) so tudi tu razvite metode na levi strani grafa. Tudi v tem primeru njihovi rezultati niso statistično značilno boljši. Podobno se tudi metodi Prec in KnnNM nista odrezali dobro. Slednja očitno sploh ni primerna za ocenjevanje zanesljivosti na teh podatkovnih množicah. Vseeno je ta graf bolj zanimiv kot zgornji, 5.1, saj je videti, da so metode razdeljene v tri skupine in sicer skupina treh na levi strani, skupina šestih na sredini in skupina štirih na desni strani. V skupini, ki se nahaja na levo, so ravno metode, ki upoštevajo, da model spada med skupinske modele. Rezultati razvitih metod niso statistično boljši, vendar je tudi tu morda razlog premajhno število podatkovnih množic. Predvsem je zanimiva razlika med najboljše uvrščeno metodo iz leve skupine in najboljšo iz sredinske skupine, ki je dolžine skoraj treh rangov.



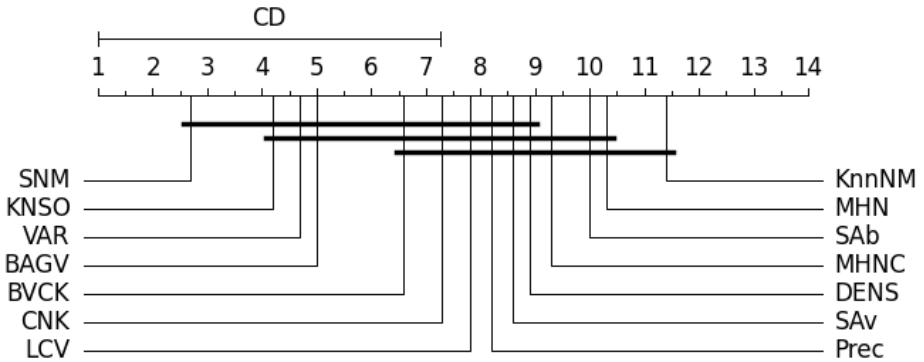
Slika 5.3: Prikaz uspešnosti metod z grafom kritične razdalje pri baggingu s 50 notranjimi modeli

Pri pregledu rezultatov prikazanih v Tabeli 5.3 smo ugotovili, da so rezultati precej podobnim rezultatom pri naključnih gozdovih (Tabela 5.1). Tako kot pri grafu rangov s kritično razdaljo je tudi tu razvidno, da se nove metode izkažejo nekoliko boljše v primerjavi z ostalimi, ker imajo vse tri statistično značilne rezultate na vsaj devetih podatkovnih množicah. Pearsonovi korelacijski koeficienti so tudi tu precej nizki, kar pomeni, da imajo ocene zanesljivosti šibko korelacijo z absolutno napako napovedi. Spremembo je mogoče opaziti tudi v deležu statistično značilnih rezultatov pri posameznih podatkovnih množicah. Pri naključnih gozdovih le dve metodi nista imeli statistično značilnega rezultata pri podatkovni množici ACHE, medtem ko je bilo takih pri modelu bagging pet. Vrstni red podatkovnih množic ni ostal enak, vendar so rezultati kljub temu podobni pri obeh skupinskih modelih.

Vrstni red metod pri testiranju s 100 notranjimi modeli je ostal približno enak (Slika 5.4). Najbolj je napredovala metoda SNM, ki je tu prepričljivejše zasedla prvo mesto. Metoda KnnNM, ki upošteva sosednost glede na notranje modele, se je odrezala tu slabše. Rezultati razvitih metod žal tudi tu niso statistično značilno boljši od obstoječih. Pri rezultatih (Tabela 5.4) je prišlo do manjših sprememb v vrstnem redu podatkovnih množic. Vrstni red metod je približno enak, najbolj je napredovala metoda MHN, ki je iz predzadnjega mesta prišla na peto, vendar so njeni rezultati prenizki, da bi napredovala tudi na grafu. Tudi tu je samo metoda SNM prišla do statistično značilnih rezultatov na vseh množicah.

Tabela 5.3 Prikaz Pearsonovih koeficientov metod na posameznih podatkovnih množicah pri skupinskem modelu bagging s 50 notranjimi modeli.

Metoda	DHFR	COX2	EDC	HIVRT	ACE	HPTP	BZR	ACHE	HIVPR	AMPH1
SNM	**0.39**	**0.27**	**0.28**	**0.42**	**0.19**	**0.33**	**0.43**	**0.23**	**0.30**	**0.28**
BAGV	**0.37**	**0.19**	**0.30**	−0.03	**0.15**	**0.25**	**0.36**	**0.28**	**0.21**	**0.26**
KNSO	**0.41**	**0.23**	**0.35**	−0.02	**0.21**	**0.35**	**0.30**	**0.30**	**0.30**	**0.34**
VAR	**0.41**	**0.23**	**0.29**	−0.05	**0.21**	**0.29**	**0.39**	**0.36**	**0.28**	**0.30**
SAv	**0.30**	**0.15**	**0.33**	**−0.39**	0.04	**0.19**	**0.20**	**0.23**	**0.14**	**0.30**
LCV	**0.34**	**0.20**	**0.20**	**0.16**	**0.18**	0.09	**0.42**	**0.29**	**0.18**	0.05
Prec	**0.23**	0.04	**0.31**	**0.23**	**0.35**	−0.04	**0.34**	−0.01	**−0.20**	**0.25**
DENS	**0.22**	**0.08**	0.09	**0.15**	**−0.19**	**0.20**	0.01	0.06	**0.33**	0.03
CNK	**0.20**	0.07	**0.36**	**0.23**	0.10	**0.24**	**0.24**	**0.21**	0.12	0.05
BVCK	**0.21**	0.08	**0.36**	**0.23**	0.11	**0.24**	**0.24**	**0.22**	0.12	0.05
SAb	**0.11**	**0.08**	−0.10	**−0.18**	**−0.18**	**0.17**	0.06	0.13	0.07	0.05
KnnNM	0.02	−0.07	**0.21**	**−0.18**	**−0.17**	−0.06	**−0.14**	−0.07	0.09	0.03
MHN	**0.15**	**0.19**	0.12	−0.01	−0.08	−0.01	−0.11	0.07	0.13	0.04
MHNC	**0.14**	**0.18**	0.11	0.04	−0.05	−0.01	−0.08	0.07	0.09	0.02



Slika 5.4: Prikaz uspešnosti metod z grafom kritične razdalje pri baggingu s 100 notranjimi modeli

Tabela 5.4 Prikaz Pearsonovih koeficientov metod na posameznih podatkovnih množicah pri skupinskem modelu bagging s 100 notranjimi modeli.

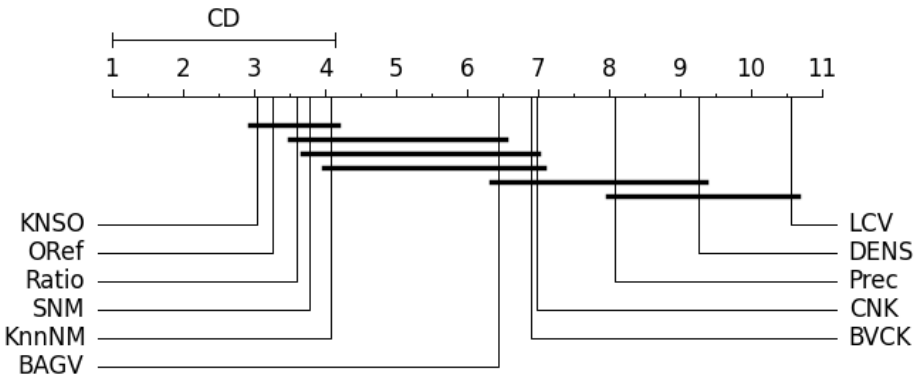
Metoda	HIVPR	BZR	DHFR	ACHE	EDC	COX2	HPTP	ACE	AMPH1	HIVRT
SNM	**0.34**	**0.40**	**0.43**	**0.24**	**0.23**	**0.21**	**0.24**	**0.19**	**0.22**	**0.38**
BAGV	**0.26**	**0.36**	**0.32**	**0.28**	**0.16**	**0.15**	**0.34**	0.11	**0.22**	0.11
KNSO	**0.26**	**0.31**	**0.41**	**0.37**	**0.26**	**0.14**	**0.36**	0.13	**0.21**	0.07
VAR	**0.26**	**0.36**	**0.41**	**0.39**	**0.22**	**0.15**	**0.39**	0.12	**0.20**	0.06
MHN	**0.19**	**−0.15**	**0.21**	**0.14**	**0.15**	**0.13**	0.01	**−0.16**	0.01	−0.01
CNK	**0.20**	**0.19**	**0.25**	**0.27**	**0.18**	0.02	**0.41**	0.12	−0.04	**0.27**
MHNC	**0.21**	**−0.15**	**0.22**	**0.14**	**0.14**	**0.13**	0.02	**−0.14**	0.04	0.06
BVCK	**0.20**	**0.19**	**0.25**	**0.27**	**0.18**	0.02	**0.41**	0.12	−0.04	**0.27**
LCV	**0.16**	**0.39**	**0.39**	**0.23**	0.11	**0.18**	**0.20**	0.01	0.04	0.10
SAb	**0.23**	**0.16**	**0.09**	**0.17**	−0.08	**0.10**	0.08	**−0.15**	0.02	0.05
DENS	**0.32**	−0.01	**0.24**	0.10	**0.30**	0.00	**0.22**	**−0.17**	0.01	−0.04
Prec	−0.05	**0.29**	**0.15**	0.06	**0.26**	0.06	0.00	**0.31**	**0.17**	0.12
SAv	**0.15**	**0.25**	**0.31**	**0.31**	0.12	0.06	**0.26**	0.04	0.12	−0.05
KnnNM	0.02	**−0.16**	−0.03	0.03	0.09	**−0.13**	**−0.13**	−0.08	−0.01	−0.07

5.2 Klasifikacijske podatkovne množice

Na klasifikacijskih podatkovnih množicah smo testirali manj metod, saj nekatere ne znajo podati ocene zanesljivosti pri diskretni napovedi. Metode smo testirali na večjem številu podatkovnih množic v primerjavi s tistimi, ki delujejo na podatkih z zvezno ciljno spremenljivko. Tu smo testirali enake skupinske modele kot na regresijskih množicah in sicer naključne gozdove in bagging. Tabele rezultatov izgledajo malenkost drugače, kot pri regresijskih podatkovnih množicah, saj smo se zaradi velikega števila podatkovnih množic odločili tabelo transponirati, torej stolpci predstavljajo metode in vrstice podatkovne množice. Tudi tu imamo enak vrstni red in sicer prva vrstica predstavlja podatkovno množico, na kateri je največ metod prišlo do statistično značilnega rezultata, prvi stolpec pa metodo, katere rezultati so bili na največ podatkovnih množicah rezultati statistično značilni. Vrednost, ki je predstavljena v celici, prikazuje normalizirano U vrednost. V nasprotju s Pearsonovim korelacijskim koeficientom tu boljši rezultat predstavljajo manjše vrednosti.

5.2.1 Naključni gozdovi

Uspešnost metod na klasifikacijskih podatkovnih množicah pri modelu naključnih gozdov kaže Slika 5.5. Nove metode se nahajajo na levi strani grafa, med katerimi je tudi metoda KNSO, ki je dosegla statistično značilno boljši rezultat v primerjavi z vsemi obstoječimi metodami razen metode ORef. Do takšnega rezultata smo lahko prišli, ker je kritična razdalja pri tem grafu precej krajša, saj smo metode, ki delujejo na klasifikacijskih problemih, testirali na bistveno več podatkovnih množicah. Metoda LCV se je pri tem modelu izkazala za najmanj uspešno.



Slika 5.5: Prikaz uspešnosti metod z grafom kritične razdalje pri naključnih gozdovih s 50 drevesi

Rezultati na posameznih podatkovnih množicah so prikazani v Tabeli 5.5, kjer je razvidno, da je precej več statistično značilnih rezultatov v primerjavi z rezultati na regresijskih podatkovnih množicah. Skrajno desno je metoda LCV, ki je dosegla statistično značilen rezultat le na dveh množicah. Metoda Ratio je prišla do statistično značilnega rezultata pri vseh podatkovnih množicah, sledile so ji metode ORef, SNM, KNSO in KnnNM z enim rezultatom, ki ni bil statistično neznačilen.

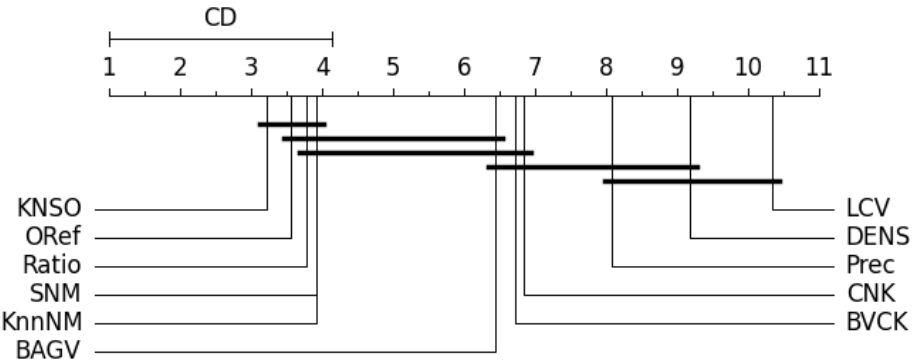
Tabela 5.5 Prikaz normaliziranih U vrednosti na posameznih podatkovnih množicah pri skupinskem modelu naključnih gozdov s 50 notranjimi modeli. Podatkovne množice so opisane v Tabeli 4.3

Metoda	Ratio	ORef	SNM	KNSO	BAGV	KnnNM	CNK	BVCK	Prec	DENS	LCV
bcwd	**0.19**	**0.10**	**0.15**	**0.16**	**0.29**	**0.14**	**0.52**	**0.52**	**0.49**	**0.47**	**0.71**
zoo	**0.05**	**0.21**	**0.12**	**0.05**	**0.22**	**0.14**	**0.02**	**0.02**	**0.13**	**0.56**	0.83
anneal	**0.26**	**0.11**	**0.12**	**0.27**	**0.33**	**0.37**	**0.58**	**0.58**	**0.54**	**0.91**	0.99
tic	**0.45**	**0.29**	**0.15**	**0.12**	**0.63**	**0.16**	**0.65**	**0.65**	**0.81**	**0.95**	0.99
iris	**0.20**	**0.17**	**0.22**	**0.23**	**0.19**	**0.26**	**0.23**	**0.23**	**0.52**	**0.58**	0.93
primary	**0.41**	**0.36**	**0.47**	**0.49**	**0.93**	**0.50**	**0.68**	**0.68**	**0.56**	**0.80**	0.97
crx	**0.45**	**0.43**	**0.51**	**0.45**	**0.62**	**0.47**	**0.75**	**0.75**	**0.70**	**0.83**	0.98
voting	**0.24**	**0.14**	**0.23**	**0.30**	**0.35**	**0.30**	**0.51**	**0.51**	**0.59**	**0.40**	0.89
bupa	**0.72**	**0.64**	**0.68**	**0.59**	**0.86**	**0.64**	**0.92**	**0.92**	**0.73**	0.96	**0.92**
shuttle	**0.05**	**0.04**	**0.06**	**0.07**	**0.70**	**0.11**	**0.26**	**0.26**	**0.37**	0.89	0.98
glass	**0.66**	**0.72**	**0.70**	**0.55**	**0.70**	**0.57**	**0.63**	**0.63**	**0.84**	0.96	0.99
monks-1	**0.08**	**0.01**	**0.00**	**0.00**	**0.43**	**0.00**	**0.92**	**0.92**	**0.13**	0.97	1.00
monks-3	**0.51**	**0.43**	**0.32**	**0.48**	**0.67**	**0.49**	**0.75**	**0.75**	0.90	**0.79**	0.91
balance-s	**0.28**	**0.29**	**0.39**	**0.23**	**0.63**	**0.33**	**0.74**	**0.74**	**0.79**	0.99	1.00
lymph	**0.38**	**0.33**	**0.39**	**0.29**	**0.76**	**0.33**	**0.57**	**0.57**	**0.86**	0.87	0.99
monks-2	**0.84**	**0.66**	**0.63**	**0.41**	**0.75**	**0.33**	**0.49**	**0.49**	**0.92**	1.00	0.99
audiology	**0.42**	0.92	**0.53**	**0.21**	**0.79**	**0.22**	**0.72**	**0.72**	**0.69**	0.92	0.95
ionosphere	**0.37**	**0.36**	**0.31**	**0.32**	**0.38**	**0.33**	**0.76**	**0.76**	0.97	0.97	0.94
post-del	**0.62**	**0.63**	**0.59**	**0.57**	0.96	**0.65**	**0.65**	**0.65**	**0.81**	0.85	0.94
horse-colic	**0.55**	**0.51**	**0.52**	**0.76**	**0.73**	**0.77**	0.95	0.95	**0.79**	**0.88**	0.96
wine	**0.08**	**0.09**	0.90	**0.06**	**0.23**	**0.08**	**0.49**	**0.50**	0.97	0.84	0.83
breast	**0.82**	**0.82**	**0.85**	0.97	**0.77**	0.98	0.94	0.94	0.94	**0.82**	0.95

Metode so se pri 100 drevesih odrezale praktično enako (Slika 5.6), česar nismo pričakovali. Tudi iz Tabele 5.6 je razvidno, da ni prišlo do večjih sprememb v rezultatih. Ena izmed razlag za tak rezultat je, da je morda pri razrednih spremenljivkah manj informacije kot pri zveznih, druga pa, da razvite metode niso znale dobro izkoristiti danih informacij.

Tabela 5.6 Prikaz normaliziranih U vrednosti na posameznih podatkovnih množicah pri skupinskem modelu naključnih gozdov s 100 drevesi.

Metoda	Ratio	ORef	SNM	KNSO	KnnNM	Prec	BAGV	CNK	BVCK	DENS	LCV
crx	**0.48**	**0.46**	**0.54**	**0.48**	**0.48**	**0.80**	**0.63**	**0.79**	**0.79**	**0.79**	**0.93**
bcwd	**0.24**	**0.12**	**0.19**	**0.15**	**0.13**	**0.51**	**0.31**	**0.50**	**0.50**	**0.50**	**0.84**
zoo	**0.04**	**0.20**	**0.06**	**0.09**	**0.16**	**0.14**	**0.14**	**0.01**	**0.01**	**0.40**	0.88
anneal	**0.21**	**0.10**	**0.08**	**0.18**	**0.19**	**0.49**	**0.35**	**0.58**	**0.58**	**0.92**	0.97
tic	**0.48**	**0.32**	**0.13**	**0.10**	**0.14**	**0.86**	**0.62**	**0.64**	**0.64**	**0.95**	0.99
iris	**0.24**	**0.22**	**0.39**	**0.31**	**0.33**	**0.65**	**0.28**	**0.21**	**0.21**	**0.58**	0.92
primary	**0.56**	**0.51**	**0.51**	**0.54**	**0.52**	**0.70**	**0.86**	**0.67**	**0.67**	**0.80**	1.00
monks-3	**0.40**	**0.39**	**0.28**	**0.44**	**0.45**	**0.80**	**0.60**	**0.67**	**0.67**	0.86	**0.85**
voting	**0.22**	**0.12**	**0.29**	**0.24**	**0.25**	**0.77**	**0.33**	**0.54**	**0.54**	**0.37**	0.99
shuttle	**0.06**	**0.04**	**0.03**	**0.04**	**0.05**	**0.39**	**0.55**	**0.24**	**0.24**	0.89	0.91
glass	**0.56**	**0.73**	**0.72**	**0.51**	**0.53**	**0.82**	**0.73**	**0.65**	**0.65**	0.93	1.00
monks-1	**0.08**	**0.02**	**0.00**	**0.00**	**0.00**	**0.05**	**0.33**	**0.90**	**0.90**	0.98	0.98
balance-s	**0.23**	**0.29**	**0.44**	**0.24**	**0.36**	**0.78**	**0.57**	**0.74**	**0.74**	0.97	0.99
monks-2	**0.82**	**0.61**	**0.57**	**0.31**	**0.31**	**0.88**	**0.70**	**0.54**	**0.54**	1.00	0.98
wine	**0.02**	**0.03**	0.97	**0.02**	**0.03**	**0.37**	**0.12**	**0.74**	**0.74**	**0.49**	0.95
audiology	**0.43**	0.96	**0.38**	**0.12**	**0.11**	**0.70**	**0.76**	**0.86**	**0.86**	0.89	0.97
ionosphere	**0.32**	**0.32**	**0.30**	**0.31**	**0.31**	0.95	**0.33**	**0.74**	**0.74**	0.97	0.89
post-del	**0.68**	**0.68**	**0.65**	**0.72**	**0.79**	**0.83**	0.92	**0.62**	**0.62**	0.98	0.97
horse-colic	**0.52**	**0.49**	**0.51**	**0.56**	**0.59**	**0.66**	**0.71**	0.97	0.97	**0.89**	0.98
lymph	**0.34**	**0.38**	**0.38**	**0.30**	**0.31**	0.88	**0.58**	**0.64**	**0.64**	0.93	0.91
breast	**0.84**	**0.82**	**0.89**	0.99	0.93	**0.91**	**0.87**	0.97	0.97	**0.82**	**0.86**
bupa	**0.73**	**0.68**	**0.75**	**0.67**	**0.71**	**0.87**	0.97	0.97	0.97	0.97	0.98



Slika 5.6: Prikaz uspešnosti metod z grafom kritične razdalje pri naključnih gozdovih s 100 drevesi

5.2.2 Bagging

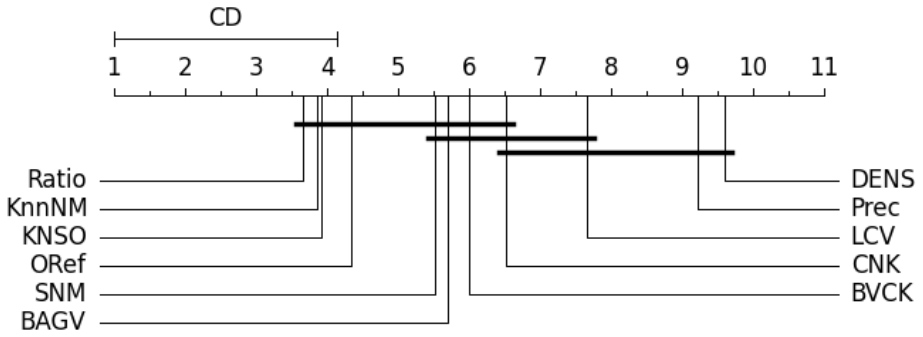
Kako so se odrezale metode pri modelu bagging, kaže Slika 5.7. Vrstni red metod je zelo podoben kot pri naključnih gozdovih, vendar je prišlo do nekaterih sprememb. Metoda LCV se je tu odrezala boljše in se v povprečju uvrščala okoli osmega mesta. Rezultati novih metod, kljub temu, da so še vedno zasedale najvišja mesta, niso bili statistično značilno boljši od obstoječih metod. Tako kot pri grafu uspešnosti metod pri modelu bagging na zveznih problemih (Slika 5.3), bi tudi tu lahko metode razdelili v skupine glede na uspešnost in sicer tiste okoli četrtega mesta, tiste okoli šestega mesta ter ostale.

Iz Tabele 5.7 je razvidno, da je le metoda BAGV prišla do statistično značilnih rezultatov na vseh podatkovnih množicah, sledile so ji Ratio, ORef, KNSO in KnnNM z enim statistično značilnim rezultatom manj. Metodi DENS in Prec sta prišli do statistično značilnega rezultata bistveno manjkrat v primerjavi z ostalimi metodami. Vrstni red podatkovnih množic je precej drugačen kot pri naključnih gozdovih, saj so vse metode dosegle statistično značilen rezultat na kar osmih podatkovnih množicah.

Podobno kot pri naključnih gozdovih, tudi tu ni prišlo do bistvenih sprememb (Slika 5.8). Najbolj vidna sprememba je mogoče rezultat metode SNM, ki se je izboljšal v povprečju za dolžino enega ranga in je tu dosegla statistično značilen rezultat na vseh množicah. Morda je metoda tu boljše izkoristila informacijo na-

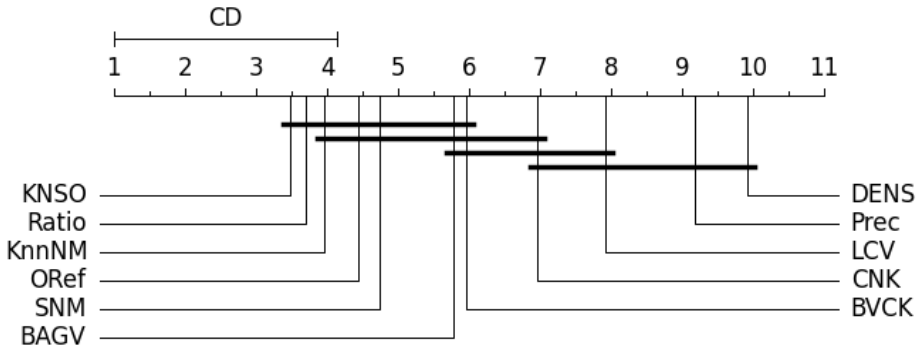
Tabela 5.7 Prikaz normaliziranih U vrednosti na posameznih podatkovnih množicah pri skupinskem modelu bagging s 50 notranjimi modeli.

Metoda	BAGV	Ratio	ORef	KNSO	KnnNM	SNM	LCV	CNK	BVCK	Prec	DENS
zoo	**0.11**	**0.03**	**0.06**	**0.04**	**0.07**	**0.11**	**0.08**	**0.05**	**0.05**	**0.10**	**0.31**
glass	**0.69**	**0.59**	**0.61**	**0.52**	**0.51**	**0.70**	**0.60**	**0.69**	**0.69**	**0.82**	**0.90**
iris	**0.18**	**0.27**	**0.27**	**0.40**	**0.27**	**0.29**	**0.37**	**0.10**	**0.10**	**0.73**	**0.60**
bcwd	**0.25**	**0.27**	**0.27**	**0.35**	**0.32**	**0.20**	**0.35**	**0.22**	**0.21**	**0.88**	**0.49**
voting	**0.42**	**0.41**	**0.41**	**0.44**	**0.43**	**0.37**	**0.41**	**0.29**	**0.26**	**0.60**	**0.39**
crx	**0.47**	**0.43**	**0.43**	**0.44**	**0.44**	**0.50**	**0.66**	**0.59**	**0.58**	**0.83**	**0.80**
audiology	**0.42**	**0.33**	**0.35**	**0.20**	**0.19**	**0.42**	**0.57**	**0.34**	**0.34**	**0.65**	0.92
anneal	**0.30**	**0.26**	**0.26**	**0.24**	**0.22**	**0.30**	**0.81**	**0.40**	**0.39**	**0.75**	0.97
primary	**0.59**	**0.49**	**0.62**	**0.61**	**0.62**	**0.61**	**0.67**	**0.66**	**0.65**	**0.62**	0.97
monks-1	**0.33**	**0.21**	**0.21**	**0.00**	**0.03**	**0.04**	**0.50**	**0.42**	**0.42**	**0.17**	0.94
monks-2	**0.79**	**0.74**	**0.74**	**0.29**	**0.27**	**0.44**	**0.60**	**0.86**	**0.86**	**0.84**	0.97
shuttle	**0.11**	**0.08**	**0.08**	**0.07**	**0.08**	**0.06**	**0.18**	**0.47**	**0.47**	0.80	0.93
tic	**0.52**	**0.30**	**0.30**	**0.11**	**0.15**	**0.17**	**0.54**	**0.34**	**0.34**	0.99	1.00
balance-s	**0.55**	**0.34**	**0.34**	**0.29**	**0.36**	**0.35**	**0.46**	**0.67**	**0.67**	0.96	0.99
lymph	**0.46**	**0.41**	**0.41**	**0.34**	**0.35**	**0.38**	**0.59**	**0.32**	**0.32**	0.90	0.89
wine	**0.27**	**0.16**	**0.16**	**0.11**	**0.13**	**0.62**	**0.41**	**0.13**	**0.12**	0.84	0.95
ionosphere	**0.33**	**0.28**	**0.28**	**0.33**	**0.26**	**0.29**	0.95	**0.46**	**0.45**	0.96	0.96
post-del	**0.74**	**0.59**	**0.59**	**0.65**	**0.74**	**0.61**	**0.69**	0.97	0.97	**0.83**	0.98
horse-colic	**0.74**	**0.64**	**0.66**	**0.70**	**0.85**	0.99	0.94	**0.87**	**0.88**	**0.84**	0.93
breast	**0.86**	**0.84**	**0.84**	**0.88**	**0.87**	**0.89**	**0.85**	0.95	0.95	0.95	**0.83**
bupa	**0.81**	**0.73**	**0.73**	**0.71**	**0.73**	**0.87**	**0.84**	0.96	0.95	0.96	0.97
monks-3	**0.56**	0.90	0.90	0.97	0.85	0.91	0.98	0.90	0.98	0.97	**0.66**



Slika 5.7: Prikaz uspešnosti metod z grafom kritične razdalje pri modelu bagging s 50 notranjimi modeli

povedi notranjih modelov in se je posledično sekundarni model bolje naučil. Iz Tabele 5.8, je razvidno, da je več metodam uspelo priti do statistično značilnega rezultata, poleg metode BAGV so sedaj statistično značilen rezultat na vseh množicah dosegle še Ratio, ORef in SNM.



Slika 5.8: Prikaz uspešnosti metod z grafom kritične razdalje pri modelu bagging s 100 notranjimi modeli

Tabela 5.8 Prikaz normaliziranih U vrednosti na posameznih podatkovnih množicah pri skupinskem modelu bagging s 100 notranjimi modeli.

Metoda	Ratio	ORef	SNM	BAGV	KNSO	KnnNM	CNK	BVCK	LCV	DENS	Prec
zoo	**0.07**	**0.12**	**0.17**	**0.21**	**0.07**	**0.09**	**0.08**	**0.08**	**0.05**	**0.38**	**0.10**
anneal	**0.25**	**0.25**	**0.21**	**0.26**	**0.21**	**0.21**	**0.43**	**0.42**	**0.84**	**0.91**	**0.76**
iris	**0.16**	**0.16**	**0.24**	**0.18**	**0.20**	**0.16**	**0.04**	**0.04**	**0.24**	**0.62**	**0.66**
primary	**0.50**	**0.64**	**0.59**	**0.58**	**0.61**	**0.60**	**0.65**	**0.65**	**0.63**	**0.88**	**0.59**
monks-1	**0.17**	**0.17**	**0.02**	**0.37**	**0.00**	**0.03**	**0.36**	**0.36**	**0.50**	**0.85**	**0.14**
voting	**0.36**	**0.36**	**0.35**	**0.36**	**0.39**	**0.37**	**0.29**	**0.28**	**0.51**	**0.50**	**0.68**
crx	**0.46**	**0.46**	**0.52**	**0.52**	**0.46**	**0.47**	**0.63**	**0.63**	**0.58**	**0.86**	**0.88**
shuttle	**0.10**	**0.10**	**0.08**	**0.11**	**0.10**	**0.12**	**0.39**	**0.39**	**0.14**	0.88	**0.81**
audiology	**0.28**	**0.29**	**0.46**	**0.39**	**0.23**	**0.22**	**0.35**	**0.35**	**0.56**	0.88	**0.74**
horse-colic	**0.65**	**0.66**	**0.88**	**0.77**	**0.81**	**0.88**	**0.90**	**0.90**	0.99	**0.92**	**0.81**
glass	**0.60**	**0.62**	**0.67**	**0.75**	**0.52**	**0.54**	**0.61**	**0.61**	**0.67**	**0.87**	0.91
breast	**0.75**	**0.75**	**0.82**	**0.79**	**0.75**	**0.75**	**0.81**	**0.81**	**0.85**	**0.75**	0.95
monks-2	**0.75**	**0.75**	**0.43**	**0.78**	**0.26**	**0.25**	**0.90**	**0.90**	**0.62**	0.96	**0.81**
bcwd	**0.24**	**0.24**	**0.22**	**0.24**	**0.29**	**0.28**	**0.17**	**0.15**	**0.34**	**0.50**	0.91
tic	**0.31**	**0.31**	**0.18**	**0.53**	**0.11**	**0.17**	**0.36**	**0.36**	**0.46**	0.99	0.95
balance-s	**0.35**	**0.35**	**0.38**	**0.59**	**0.23**	**0.32**	**0.66**	**0.65**	**0.49**	0.99	0.97
lymph	**0.39**	**0.38**	**0.42**	**0.47**	**0.28**	**0.33**	**0.48**	**0.48**	**0.57**	0.94	0.92
wine	**0.21**	**0.21**	**0.65**	**0.20**	**0.14**	**0.16**	**0.12**	**0.12**	**0.59**	0.89	0.78
ionosphere	**0.45**	**0.45**	**0.48**	**0.45**	**0.42**	**0.44**	**0.52**	**0.52**	0.94	0.97	0.99
bupa	**0.75**	**0.75**	**0.84**	**0.88**	**0.76**	**0.77**	**0.92**	**0.92**	0.97	0.96	0.93
post-del	**0.71**	**0.71**	**0.68**	**0.73**	**0.80**	**0.80**	0.98	0.97	**0.77**	0.95	0.85
monks-3	**0.72**	**0.72**	**0.58**	**0.62**	0.95	0.90	0.92	0.94	0.79	0.84	0.95

Poglavje 6

Sklepne ugotovitve

Kljub temu, da v nobenem izmed rezultatov nismo dosegli statistično značilno boljši rezultat od obstoječih metod, na vprašanje ali lahko bolje ocenimo zanesljivost v kolikor vemo, da je napoved podal skupinski klasifikator, je naš odgovor, da se najverjetneje da. Pri rezultatih na zveznih podatkovnih množicah nobena izmed razvitih metod ni bila statistično značilno boljša od obstoječih, a so se metode razvite za regresijske probleme kljub temu konsistentno uvrščale na prva tri mesta. Predvsem prepričljiv rezultat je dosegla metoda Stacking notranjih modelov (SNM), ki je imela vedno najnižji povprečen rang, kar je zelo malo verjetno posledica naključja. Rezultati na klasifikacijskih podatkovnih množicah povsod prikazujejo nekatere razvite metode na levi strani grafa rangov s kritično razdaljo, kar pomeni, da so metode dosegale v povprečju boljše rezultate. Metoda KNSO, ki je bila razvita za ocenjevanje napovedi pri regresijskih problemih, se je na klasifikacijskih množicah odrezala bolje kot metode razvite primarno za ocenjevanje zanesljivosti klasifikacijskih napovedi. Potrebno je omeniti, da kljub temu, da je metoda ORef v večini grafov bila na levi strani, še ne moremo sklepati, da so metode na desni strani neuporabne. Graf rangov s kritično razdaljo prikazuje povprečen rang na vseh podatkovnih množicah, torej se je lahko katera izmed metod odrezala na kakšni podatkovni množici boljše kot metoda ORef, vendar je na koncu še vedno dosegla v povprečju slabši rang. Za take metode lahko rečemo, da so uporabne na specifičnih množicah. To velja tudi za obstoječe metode, ki so v večini na desni strani, saj se je, na primer, metoda CNK pri naključnih gozdovih na podatkovni množici zoo odrezala najboljše.

Menimo, da bi bilo vredno preizkusiti razvite metode, zlasti SNM, na več regresijskih podatkovnih množicah, kjer bi z manjšo kritično razdaljo morda prišli do statistično značilno boljšega rezultata. Poleg tega smo se v delu osredotočili le na metode, ki podajo oceno zanesljivosti v numerični obliki in bi vsekakor bilo vredno poizkusiti razviti kakšno metodo, ki kot oceno zanesljivosti vrne denimo napovedni interval.

Literatura

[1] Siddhartha Bhattacharyya. Confidence in predictions from random tree ensembles. *Knowledge and information systems*, 35(2):391–410, 2013.

[2] Zoran Bosnić and Igor Kononenko. Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering*, 67(3):504–516, 2008.

[3] Zoran Bosnić and Igor Kononenko. Estimation of individual prediction reliability using the local sensitivity analysis. *Applied intelligence*, 29(3):187–203, 2008.

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

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

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

[7] Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. `http://jmlr.org/papers/v14/demsar13a.html`.

[8] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. `http://www.scipy.org/`.

[9] Rok Močnik. Metode za oceno zanesljivosti napovedi učinkov malih molekul. Diplomsko delo, Univerza v Ljubljani, 2011.

[10] Kristine Monteith and Tony Martinez. Using multiple measures to predict confidence in instance classification. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

[11] David A Nix and Andreas S Weigend. Learning local error bars for nonlinear regression. *Advances in neural information processing systems*, pages 489–496, 1995.

[12] Darko Pevec, Erik Štrumbelj, and Igor Kononenko. Evaluating reliability of single classifications of neural networks. In *Adaptive and Natural Computing Algorithms*, pages 22–30. Springer, 2011.

[13] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, volume 2, pages 722–726, 1999.

[14] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *The Journal of Machine Learning Research*, 9:371–421, 2008.

[15] Marko Toplak and Matija Polajnar. Orange reliability documentation, 2014. `http://www.pythonhosted.org/Orange-Reliability/`.

[16] Marko Toplak, Rok Mocnik, Matija Polajnar, Zoran Bosnic, Lars Carlsson, Catrin Hasselgren, Janez Demsar, Scott Boyer, Blaz Zupan, and Jonna Stalring. Assessment of machine learning reliability methods for quantifying the applicability domain of qsar regression models. *Journal of chemical information and modeling*, 54(2):431–441, 2014.

[17] Huazhen Wang, Chengde Lin, Fan Yang, and Xueqin Hu. Hedged predictions for traditional chinese chronic gastritis diagnosis with confidence machine. *Computers in biology and medicine*, 39(5):425–432, 2009.

[18] Achilleas Zapranis and Efstratios Livanis. Prediction intervals for neural network models. In *Proceedings of the 9th WSEAS International Conference on*

Computers, page 76. World Scientific and Engineering Academy and Society (WSEAS), 2005.