Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Iztok Oder

Napovedovanje Alzheimerjeve bolezni z zlivanjem podatkov

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: izr. prof. dr. Zoran Bosnić

Ljubljana, 2018

Zahvala

Zahvaliti se želim prof. Zoranu Bosniću za vse nasvete in strokovno pomoč pri izdelavi magistrske naloge, predvsem pa za izjemno hitro odzivnost na moja elektronska sporočila z vprašanji. Prav tako bi se Vam rad zahvalil za neverjetno razumevanje in potrpežljivost pri mojem počasnem pisanju. Hvala tudi za priložnost sodelovanja pri pisanju znanstvenega članka.

Posebej bi se rad zahvalil vsem sošolcem, s katerimi sem preživel najlepša leta študija. Hvala Milutin, Ernest, Žiga, Matic, Nina, Darja, Eva, Matevž in Jaka za vso motivacijo, zabavo in trenutke, ki smo jih preživeli skupaj.

Zahvala gre tudi mojim staršem, ki sta me podpirala v času študija, in vsem prijateljem za vse nepozabne dogodivščine.

Iztok Oder, 2018

"I have not failed. I've just found 10000 ways that won't work."

— Thomas A. Edison

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
MRI	magnetic resonance imaging	slikanje z magnetno resonanco
PET	positron emission tomography	pozitronska emisijska tomografija
EEG	electroencephalography	elektroencefalografija
SVM	support vector machine	metoda podpornih vektorjev
LR	logistic regression	logistična regresija
DT	decision tree	odločitveno drevo
KNN	k-nearest neighbours	k-najbližjih sosedov
NB	naive Bayes	naivni Bayes
RF	random forest	naključni gozd
NN	neural network	nevronska mreža
AUC	area under curve	površina pod krivuljo
ANCOVA	analysis of covariance	analiza kovariance
CD	critical difference	kritična razdalja

Povzetek

Naslov: Napovedovanje Alzheimerjeve bolezni z zlivanjem podatkov

Alzheimerjeva bolezen je vse bolj pereč problem v modernem svetu, saj se svetovno prebivalstvo stara. Ta zahrbtna nevrodegenerativna bolezen povzroča demenco in vpliva na vsakodnevno življenje tako obolelega kot njegovih oskrbnikov. Diagnoza bolezni je trenutno mogoča šele po smrti osebe, ki trpi za to boleznijo. Raziskovalci se zato trudijo z različnimi tipi podatkov in metodami izboljšati modele za dovolj zgodnje razlikovanje med zdravimi in obolelimi osebami.

V magistrski nalogi obravnavamo napovedovanje Alzheimerjeve bolezni z uporabo zlivanja podatkov. V ta namen predlagamo metodo zgodnje integracije atributov. Metoda na skupnih atributih dveh množic nauči modele za nadzorovano učenje in z njimi napove manjkajoče atribute iz druge množice atributom iz prve množice. Implementirano metodo stestiramo na več množicah, ki vsebujejo podatke različnih modalnosti, tako da med seboj primerjamo večje število modelov, zgrajenih na osnovnih in zlitih množicah. Obravnavane implementacije metod značilno ne izboljšajo modelov. V nekaterih primerih pa je vseeno mogoče opaziti izboljšanje napovedi in tako potencial zlivanja podatkov za zvišanje klasifikacijske točnosti.

Ključne besede

Alzheimerjeva bolezen, strojno učenje, zlivanje podatkov, napovedovanje, klasifikacija

Abstract

Title: Predicting the Alzheimer's disease using data fusion

Alzheimer's disease is an increasingly pressing problem in the modern world, as the world's population is aging. This insidious neurodegenerative disease causes dementia and affects the daily life of both – the diseased and their caregivers. Diagnosing this disease is currently possible only after the death of the affected person. Researchers are striving to develop a method for predicting this disease using different types of data.

In this master's thesis we are trying to improve prediction of Alzheimer's disease using data fusion. To this end, we propose an early attribute integration method. Our method takes common attributes of two sets and builds models to extend the attributes from original set with the missing attributes of the other set. We test the our implementation of the proposed method on multiple sets containing data of various modalities. We compare classification models built on the fused datasets and on the original datasets. Our implementation of the method does not significantly improve the models. However, in some cases, improvement is noticeable and can nevertheless indicate the potential of data fusion for increasing the classification accuracy.

Keywords

Alzheimer's disease, machine learning, data fusion, prediction, classification

Poglavje 1

Uvod

Demenca je sindrom, ki se pojavi zaradi bolezni možganov. Pri tem pride do zmanjšanja oziroma pojemanja mentalnih in intelektualnih funkcij. Med njih spadajo mišljenje, spomin, presoja, zaznava, orientacija, sklepanje, sposobnost učenja, sposobnost reševanja problemov, računanje itd. [1] Okoli 47 milijonov ljudi po vsem svetu trpi za demenco in vsako leto se ta številka poveča za 9.9 milijonov. Svetovna zdravstvena organizacija pričakuje do leta 2030 porast obolelih na 75 milijonov [2].

Demenca se lahko razvije pri komurkoli. Pri mladih ljudeh se lahko nenadoma razvije, kadar pride do uničenja možganskih celic zaradi hude poškodbe ali bolezni. Ponavadi se demenca počasi razvija pri ljudeh, starejših od 60 let, vendar ni normalen del staranja. Staranje povzroča spremembe pri vseh ljudeh. Pojavljati se začnejo vrzeli v spominu, težje postane opravljanje več stvari hkrati, učenje novih stvari zahteva več časa itd. Te spremembe vplivajo na izgubo razuma do te mere, da oseba ni več zmožna normalnega delovanja in vsakodnevnega življenja [3].

Vzrokov za nastanek demence je več. Lahko se pojavi po hudi možganski poškodbi, srčnem ali možganskem infarktu ali pa je posledica ene izmed več kot 200 znanih bolezni. Med njih spadajo med drugim tudi Alzheimerjeva bolezen, bolezen z Lewyjevimi telesci, Parkinsonova bolezen, AIDS itd. [1] Najpogostejši razlog za pojavitev demence je Alzheimerjeva bolezen. Svetovna zdravstvena organizacija pripisuje med 60 in 70 odstotkov primerov demence Alzheimerjevi bolezni [2].

Alzheimerjeva bolezen je nevrodegenerativna bolezen možganov. Gre za zelo zahrbtno bolezen, saj povzroča možganske poškodbe kar nekaj let, preden se pojavijo kakršnikoli simptomi čustvene, fizične ali kognitivne narave. Točni vzroki za nastanek bolezni še niso znani, vendar so raziskovalci odkrili precej dejavnikov tveganja, kot so: starost, ženski spol, genetika, visok krvni tlak, visok holesterol, diabetes, debelost in motnje spanja [4]. Postavljanje diagnoze je zaradi tega izjemno težko. Vztrajno napredovanje demence in z

Slika 1.1: *Leva slika prikazuje normalne možgane. Desna prikazuje atrofijo možganov, ki je posledica Alzheimerjeve bolezni. Iz slike je tudi razvidno zmanjšanje volumna in povečanje razmaka med možganskimi režnji. Vir slike: Alzheimer's association*[1].

njo povezanih težav, kot je izguba spomina, nakazuje pri starejših ljudeh na prisotnost Alzheimerjeve bolezni. Zdravniki v teh primerih opravijo vrsto preiskav in pogovorov z bolnikom in njegovimi svojci. Za bolezen je značilna akumulacija nenormalnih beljakovin v obliki nevrofibrilarnih pentelj (angl. neurofibrillary tangles) in senilnih plak (angl. amyloid plaques). Akumulacija škodljivih beljakovin povzroča odmiranje nevronskih celic in posledično atrofijo možganov. Razliko med zdravimi in poškodovanimi možgani prikazuje slika 1.1. Identifikacija teh beljakovin je mogoča le z analizo možganskega tkiva. Posledično se lahko pravilna diagnoza postavi šele po smrti [5].

Alzheimerjeva bolezen je trenutno še neozdravljiva, vendar obstajajo zdravila, ki pomagajo prizatetim ljudem in za nekaj let upočasnijo razvoj bolezni. Poleg zdravil obstaja tudi precej terapevtskih pristopov, ki zaustavljajo hitro napredovanje bolezni. Čimprejšnja diagnoza ljudi s kognitivno motnjo je zato zelo pomembna, saj jim to omogoči boljšo izbiro nege in tako podaljša samostojnost in posledično kakovost življenja. Bolezen vpliva tudi na oskrbovalce, ki so najpogosteje svojci. Z zgodnjim odkritjem se tako lahko bolje pripravijo na prihodnost. To vključuje med drugim tudi pogovor o financah, medicinski oskrbi, različnih terapijah itd. [5, 6]

Napredki v medicinskem slikanju, identifikaciji biomarkerjev v krvni plazmi in napreden razvoj nevropsiholoških testov omogočajo uporabo metod strojnega učenja za iskanje prediktorjev Alzheimerjeve bolezni in njeno napovedovanje. V tej magistrski nalogi se bomo osredotočili na zlivanje različnih podatkovnih množic z namenom izboljšanja napovedne točnosti Alzheimerjeve bolezni.

V nadaljevanju najprej naredimo pregled sorodnih del in opišemo dosežke na področju analize Alzheimerjeve bolezni. V tretjem poglavju so podrobno opisane vse uporabljene metode strojnega učenja. V četrtem poglavju predstavimo našo metodo zlivanja podat-

[1] `https://www.alz.org/braintour/alzheimers_changes.asp`

kov. V petem poglavju predstavimo postopek učenja, uporabljene podatke in metode za ocenjevanje uspešnosti. V šestem poglavju predstavimo rezultate in jih ocenimo.

Poglavje 2

Pregled literature

V literaturi se Alzheimerjeva bolezen (AB) pojavlja od leta 1907, ko jo je prvi opisal dr. Alois Alzheimer. Od takrat na tem področju aktivno raziskujemo različne možnosti diagnosticiranja bolezni. Veliko poudarka je predvsem na področjih algoritmov za napovedovanje stadija bolezni in določanja prediktorjev bolezni. Stadij bolezni je omejen na tri skupine, kamor delimo osebe glede na stopnjo demence. Kontrolna skupina ne izkazuje nobenih znakov demence. Osebe z blažjo kognitivno motnjo kažejo prisotnost demence, vendar so zmožne vsakodnevnega življenja. V zadnjo skupino spadajo osebe, ki jasno izkazujejo znake bolezni.

V zadnjih 20 letih so napredki v razvoju medicinskih naprav omogočili zajemanje različnih tipov podatkov in tako odprli nove možnosti za analizo bolezni. Med njih spadajo magnetno-resonančne slike (MRI), slike pozitronske emisijske tomografije (PET), signali elektroencefalografa (EEG), biomarkerji iz krvi in cerebrospinalne tekočine, rezultati kliničnih in kognitivnih testov in demografski podatki.

V tem poglavju predstavimo dosedanje dosežke na področju raziskovanja Alzheimerjeve bolezni. Poglavje je razdeljeno na podpoglavja glede na podatke, ki so jih raziskovalci primarno uporabili v člankih.

2.1 Uporaba medicinskih slik

Večina raziskovalcev uporablja podatke samo ene modalnosti v kombinaciji že uveljavljenimi metodami strojnega učenja, kot so na primer nevronske mreže in metoda podpornih vektorjev (SVM). Najpogosteje se za napovedovanje AB uporabljajo slike PET ali slike MRI, iz katerih je mogoče izluščiti različne vrste atributov. Surovi podatki teh slik predstavljajo jakosti 3D množice vokslov, ki jih je možno uporabiti kot atribute. Tak vektor atri-

butov je poln šuma in je za klasifikacijo uporaben šele po uporabi algoritma za zmanjšanje dimenzionalnosti. Obstajajo pristopi, kjer lahko z algoritmi računalniškega vida izluščimo ogromno množico različnih opisnih atributov iz slik, kot so debelina možganske skorje, volumni in ukrivljenost možganskih regij in podobne atribute posameznih delov možganov [7, 8, 9].

Chu et al. so v članku [10] raziskovali vpliv različnih metod za izbiranje atributov. Preizkusili so štiri različne metode. Prva je voksle iz slik sive možganovine razdelila na 27 različnih regij, ki so jih dodatno zmanjšali na 9 z uporabo rezultatov iz drugih študij. Druga metoda je voksle oseb z AB primerjala z voksli zdravih oseb z uporabo t-testa. Tretja je bila kombinacija prvih dveh. Najprej so voksle primerjali z uporabo t-testa, nato so izračunali povprečno t-vrednost v posamezni regiji in na koncu razvrstili regije glede na povprečne t-vrednosti. Najboljših n so uporabili za klasifikacijo. Četrta metoda je uporabljala rekurzivno eliminacijo atributov z uporabo predelane SVM [11]. Za klasifikacijo so uporabili SVM. S prvo in tretjo metodo so dosegli klasifikacijske točnosti okoli 85%. Ti metodi sta prinesli značilno boljše rezultate kot klasifikacija brez izbire atributov.

Liu et al. [12] so predlagali uporabo ansambla, ki je uporabljal šibke klasifikatorje z redko predstavitvijo podatkov (angl. Sparse representation-based classifier). Ta klasifikator v fazi učenja z uporabo vseh podatkov zgradi slovar. V fazi napovedovanja testnega primera v slovarju poišče njegovo redko predstavitev. Zakodira ga kot linearno kombinacijo vseh učnih primerov in z uporabo normalizacije L1 evalvira rekonstrukcijsko napako za vsak razred. Razred z najmanjšo napako uporabi za napoved. Vhod v vsak šibki klasifikator je bila naključna podmnožica zaplat vokslov vnaprej določenih velikosti ($3\times3\times3$, $5\times5\times5$, ...). Predlagano metodo so primerjali s SVM z linearnim jedrom. Z naraščajočim številom vokslov je točnost SVM padala, medtem ko je točnost ansambla rastla. Ansambel je dosegel točnost 91%. Analizirali so tudi posamezne točnosti šibkih klasifikatorjev. Ugotovili so, da so šibki klasifikatorji z visokimi točnostmi uporabljali zaplate iz možganskih regij, ki se jih pogosto povezuje z Alzheimerjevo boleznijo. Med njih spadajo med drugim hipokampus, entorinalna skorja, amigdala, ...

Napovedovanja Alzheimerjeve bolezni so se v [13] lotili z uporabo algoritma LPboosting (angl. Linear Programming Boosting). Z agregacijo napovedi šibkih klasifikatorjev algoritem poskuša izboljšati napovedno točnost celotnega modela. Avtorji so za šibke klasifikatorje uporabili kar vrednosti posameznih vokslov, transformiranih z uporabo prilagojene sigmoidne funkcije. Razdelitvena ravnina, ki jo je izračunal algoritem, je bila odvisna od napake posameznega šibkega klasifikatorja. Za zmanšanje pretiranega prilagajanja učnim primerom je algoritem uporabljal regularizacijo L1. Algoritem so uporabili na vokslih slik MRI, kjer je dosegel točnost 82% in PET slik, kjer je bila točnost 80%. Tako kot v [12] je tu analiza uteži posameznih šibkih klasifikatorjev pokazala, da so klasifikatorji z visokimi utežmi pripadali vokslom iz možganskih regij, ki so jih tudi druge študije

označile za prediktorje bolezni. Poleg tega so za nekatere primere slik ugotovili, da je več kot 65% šibkih klasifikatorjev narobe ocenilo napoved razreda, kar nakazuje na možnost, da razred že na začetku ni bil pravilno dodeljen.

Avtorji članka [14] so se napovedovanja AB lotili z uporabo konvolucijske nevronske mreže (CNN) na surovih slikah MRI. Predlagan klasifikator uporablja spodnje nivoje mreže za generalizacijo vektorja vokslov v stisnjen vektor atributov, ki jih lahko višji nivoji uporabijo za učenje binarne ali večrazredne klasifikacije. Točnost napovedi AB je z uporabo prečnega preverjanja dosegla kar 97%.

V članku [15] so uporabili PCA (angl. Principal Component Analysis) za zmanjšanje števila atributov pridobljenih iz vokslov PET slik. Pri tem niso uporabili klasične metode za predstavitev podprostora za izbiro baznih vektorjev glede na zajeto varianco, temveč so jih izbrali glede na izračunan Fisherjev količnik, za katerega so uporabili tudi informacijo o razredu. Izračunane atribute so uporabili na napoved AB z nevronskimi mrežami in SVM z linearnim in kvadratnim jedrom. Dosegli so klasifikacijsko točnost okoli 90% in senzitivnost okoli 91%.

Kippenhan et al. so v [16] voksle iz PET slik razdelili v 67 regij in za vsako izračunali povprečno aktivnost metabolizma. Z dobljenimi podatki so analizirali napovedno moč LDA (angl. Linear Discriminant Analysis) in nevronskih mrež. Nevronske mreže so se izkazale za boljšo izbiro kot LDA. Rezultate so primerjali z napovedmi eksperta na tem področju in ugotovili, da so primerljivi z nevronskimi mrežami.

Za klasifikacijo so v [17] prav tako uporabili LDA. Vhodne podatke v algoritem so izračunali iz slik MRI in znižali dimenzionalnost s PCA. Vhodni podatki v PCA so bile debeline možganske skorje, izračunane v vsakem oglišču 3D rekonsktrukcije možganske skorje, ki so jih transformirali v frekvenčno domeno in odstranili visokofrekvenčne komponente. Dosegli so specifičnost 82% in senzitivnost 93%. Najbolj diskriminativna oglišča so bila oglišča regij entorinalne in prefrontalne skorje.

Uporaba longitudinalnih podatkov (podatki zajeti ob različnih pregledih skozi čas) lahko omogoči doprinos novih informacij o posamezni osebi. To so pokazali v [18], kjer so z uporabo procesiranih atributov iz longitudinalnih slik MRI značilno izboljšali napovedno točnost klasifikatorja SVM. Za vsako osebo so imeli po dve sliki – iz prvega pregleda in iz ponovnega pregleda po 12 mesecih. Iz njih so za 83 regij izračunali povprečne intezitete vokslov. Napovedni točnosti za poskuse z atributi iz posameznih pregledov so bile značilno slabše kot za poskus, kjer so bili upoštevani atributi obeh pregledov hkrati. Točnost napovedi Alzheimerjeve bolezni tega poskusa je bila 88%, senzitivnost 83% in specifičnost 94%.

Huang et al. [19] so prav tako eksperimentirali z longitudinalnimi slikami MRI. Namesto gradnje enega klasifikatorja z optimalno podmnožico atributov so predlagali hierarhični klasifikator, ki uporablja večje število preprostih klasifikatorjev, razdeljenih v tri nivoje.

Pomembni lastnosti predlaganega klasifikatorja sta zmnožnost naslovitve problema previsoke dimenzionalnosti podatkov in vključitev in uporaba prostorskih informacij. Prvi nivo je zgradil klasifikatorje na surovih slikah MRI – intenzitetah vokslov. Na drugem nivoju je vsak klasifikator uporabil zaplato izhodov iz prejšnjega nivoja. Zadnji nivo je združil vse napovedi v eno samo. Predlagana metoda je imela točnost 79% in je bila boljša od posameznega klasifikatorja.

2.2 Uporaba podatkov o krvnih proteinih

V literaturi se vse pogosteje pojavlja uporaba podatkov o koncentracijah proteinov iz krvne plazme. Zaradi lažjega odvzema vzorca in manjše invazivnosti so metode, ki takšne podatke uporabljajo, seveda optimalnejše. Doecke et al. [20] so za napovedovanje bolezni uporabili podatke z meritvami približno 180 različnih proteinov. Za izločitev nerelevantnih atributov so opravili kompleksno analizo. Iz učnega dela podatkov so izbrali 70% primerov in na njih uporabili 4 različne metode za izbiro atributov. Ta postopek so ponovili 100-krat. Atributi, ki jih je vsaka metoda izbrala v več kot 50% poskusov, so bili uporabljeni za nadaljnje analize. Najpogosteje izbrani proteini se pojavljajo tudi v drugih raziskavah. Poleg tega so odkrili tudi protein (angl. Carcinoembryonic antigen), ki ga do takrat še nihče ni povezal z Alzheimerjevo boleznijo. Na podlagi izbranih atributov so tudi napovedali bolezen z metodami naivni Bayes, naključni gozdovi in SVM. Senzitivnost in specifičnost so dosegli 85%, AUC pa 0.89.

V članku [21] so zgornjo raziskavo še dodatno poglobili. Iz skupine 120 proteinov so s statističnim algoritmom določili 18 takih, ki so značilno ločili primere z Alzheimerjevo boleznijo od kontrolnih primerov. Uporabili so algoritem SAM (angl. Significance Analysis of Microarrays), ki vsak protein testira z uporabo prilagojenega t-testa. Testiranje se ponovi večkrat, vsakič pa se vrednosti posameznega proteina permutirajo med ciljnimi skupinami. Rezultate vseh testov se primerja in določi, ali je permutiranje vplivalo nanje. Za značilne proteine so poiskali tudi biološko razlago. Z uporabo orodja za določanje genskih omrežij so identificirali dve neodvisni regulatorni omrežji, ki povezujeta proteine in določata njihovo medsebojno odvisnost. Disfunkcija posamezne poti v omrežju vpliva na prisotnost proteinov v krvi, kar pa lahko posledično nakazuje na prisotnost bolezni.

Llano et al. [22] so primerjali več metod za izbiro atributov v kombinaciji s klasifikatorji. Analizirali so napovedno moč podatkov o koncentracijah proteinov za napovedovanje AB in kontrolnega razreda z uporabo testa ANCOVA in t-testa. Na značilnih prediktorjih so dodatno uporabili eno izmed metod za izbiro atributov: naključni gozd, metoda delnih najmanjših kvadratov, bagging in simulirano ohlajanje. Na optimalni podmnožici so nato zgradili klasifikacijski model z enim od naslednjih algoritmov: naključni gozd, dia-

gonalna linearna diskriminantna analiza, SVM, umetna nevronska mreža, metoda delnih najmanjših kvadratov, bagging in k-najbližjih sosedov. Najvišjo klasifikacijsko točnost so dosegli z uporabo naključnega gozda za gradnjo modela.

2.3 Uporaba drugih vrst podatkov

Raziskovalci uporabljajo tudi druge vrste podatkov. Med njih spadajo signali EEG, funkcijske slike MRI, rezultati kognitivnih testov in biomarkerji, kot je cerebrospinalna tekočina.

Redkeje lahko zasledimo uporabo signalov EEG. Signali v surovi obliki niso primerni za obdelavo, saj vsebujejo preveč šuma. V [23] so izmerili signale EEG na 19 lokacijah na glavi in jih s Fourierjevo transformacijo transformirali v frekvenčno domeno. Za vsako od 19 transformacij so izračunali povprečno magnitudo signala za 5 frekvenčnih območij. Iz dobljenih rezultatov so dodatno izračunali nelinearne atribute. Vse obdelane atribute so združili in jih uporabili kot vhod za algoritme strojnega učenja: diskriminantno analizo, 1-NN in nevronske mreže. Najbolje se je odrezala nevronska mreža s točnostjo 92%, senzitivnostjo 86% in specifičnostjo 96%. Ugotovili so tudi, da se frekvenčna območja precej spremenijo pri osebah z boleznijo. Aktivnost je značilno padla predvsem v območjih od 8.0-12.5 Hz in od 13.0-18.0 Hz.

Challis et al. [24] so primerjali uspešnost logistične regresije Bayesovo-Gaussovih procesov (angl. Bayesian Gaussian process logistic regression) z linearnimi in nelinearnimi kovariančnimi funkcijami z SVM modelom. Vhodni podatki so bili funkcijske slike MRI. Zgrajeni modeli so bili zelo primerljivi med seboj, saj so vsi dosegali napovedne točnosti okoli 80% pri razlikovanju blažje kognitivne motnje od kontrol.

Galili et al. [25] so uporabili popolnoma nov pristop k dodeljevanju diagnoze primerom. Njihovo idejo je gnalo nezaupanje v originalno postavitev diagnoze. Z uporabo kliničnih podatkov in originalne diagnoze so razdelili primere v 10 novih kategorij bolezni. Za to so uporabili metodo k voditeljev. Preostale podatke (biomarkerje) in nove ciljne spremenljivke so uporabili za izgradnjo klasifikacijskega drevesa z uporabo algoritma CART (angl. Classification And Regression Trees).

Tierney et al. [26] so raziskali možnost uporabe nevropsiholoških testov za napoved nastopa Alzheimerjeve bolezni. Z uporabo testa ANCOVA so analizirali razlike med osebami z AB in kontrolnimi pacienti. Ugotovili so, da so rezultati testov MMSE (angl. Mini-Mental State Examination) in DRS (angl. Dementia Rating Scale) značilno drugačni za ti dve skupini. Prav tako so uporabili logistično regresijo za nadaljnjo analizo napovedne moči testov. Najprej so izračunali interkorelacijsko matriko za vse atribute in izločili tiste z visoko korelacijo. Preostale atribute so uporabili za izgradnjo modela, katerega točnost

je bila 89%.

Gomar et al. [27] in Ewers et al. [28] so preiskovali zmožnost biomarkerjev in kognitivnih testov napovedovanja sprememb iz blage kognitivne motnje v Alzheimerjevo bolezen. Gomar et al. [27] so uporabili test ANOVA in t-test za določitev značilnih atributov. Te atribute so nato uporabili za izgradnjo modela z logistično regresijo. Napovedna točnost je bila 72%. Ewers et al. [28] so najprej določili najboljši napovedni model za diskriminacijo med AB in kotrolami. Najboljši model je bil uporabljen za napoved prehoda iz blažje kognitivne motnje v AB. Z uporabo Coxove regresije so modelirali čas do prehoda med stadijema bolezni. Med najboljšimi prediktorji sta bila tudi entorinalna skorja in hipokampus.

2.4 Uporaba zlivanja podatkov

Integracija podatkov različnih modalnosti zahteva bolj kompleksen algoritem za uspešen izkoristek vseh dopolnilnih informacij, ki jih modalnosti vsebujejo. Raziskovalci se tega lotevajo z različnimi metodami zlivanja, ki jih lahko uvrstimo v tri skupine: metode zgodnje, vmesne in pozne integracije. Pri zgodnji integraciji gre za preprosto konkatenacijo podatkov po nekem ključu. Pri vmesni integraciji so podatki posameznih modalnosti vhod v algoritem strojnega učenja, ki izlušči informacije in jih uporabi za učenje [29]. Pri pozni integraciji se združuje napovedi algoritmov strojnega učenja, ki jih učimo na podatkih posameznih modalnosti, v končno napoved [30]. Zlivanje podatkov lahko razširimo tudi na podatkovne množice iz različnih virov. Pri tem množice ne vsebujejo nekega eksplicitnega ključa, po katerem bi jih lahko združili. V literaturi take uporabe fuzije nismo zasledili.

Zhang et al. so v [29] uporabili metodi zgodnje in vmesne integracije. Za to so uporabili podatke treh modalnosti: volumne regij iz slik MRI, povprečne intenzitete vokslov regij iz slik PET in biomarkerje iz cerebrospinalne tekočine. Prispevke posameznih modalnosti so uporabili v večjedrnem SVM. Za vsako modalnost so izračunali svoje jedro, njihovo uteženo vsoto pa uporabili v SVM. To metodo so primerjali s SVM, ki je uporabljal podatke ene same modalnosti, in s SVM, ki je uporabljal podatke vseh modalnosti združene v eno tabelo. Večjedrni SVM je dosegal višje napovedne točnosti od vseh ostalih metod. Večjedrni SVM in SVM, ki je uporabljal združene podatke, sta bili značilno boljša od metod, ki so uporabljale podatke ene modalnosti.

Tudi R. Gray et al. so v [30] uporabili vmesno integracijo za zlivanje podatkov različnih modalnosti (MRI, PET, cerebrospinalno tekočino in gene). Za vsako modalnost so zgradili model z uporabo nakjučnega gozda. Z uporabo vsakega od zgrajenih modelov so izračunali podobnosti med primeri. Tako so dobili matrike podobnosti za vsako modalnost, katerih linearno kombinacijo so uporabili za izračun optimalne raznolike predstavitve (angl. ma-

nifold representation). Na novo dobljeni predstavitvi primerov so zgradili končni model za napovedovanje bolezni. Večmodalni model se je izkazal značilno boljšega od modelov, ki so uporabljali podatke ene modalnosti.

Suk et al. [31] so uporabili slike MRI in PET za napovedovanje bolezni. Iz njih so izluščili skupine vokslov, ki so se izkazale za značilne pri ločevanju med primeri Alzheimerjeve bolezni in kontrolnimi pacienti. Iz teh skupin so z uporabo večmodalne DBM (angl. Deep Bolzman Machine) izluščili atribute posameznih modalnosti in jih na najnižjem nivoju te nevronske mreže združili. Dobljene atribute so uporabili v ansamblu iz [12]. Pri tem so kot šibek klasifikator uporabili SVM. Predprocesiranje skupin je izboljšalo napovedno točnost ansambla v vseh poskusih – napovedovanje z uporabo posameznih modalnosti in napovedovanje z zlivanjem obeh modalnosti. Prav tako je zlivanje izboljšalo napovedno točnost v primerjavi z točnostjo napovedi s podatki ene same modalnosti. Napovedna točnost je bila 95%. Analiza možganskih regij, ki so prispevale k pravilni klasifikaciji, je potrdila rezultate drugih študij. Med najbolj pomembne so se ponovno uvrstile hipokampus, entorinalna skorja in amigdala.

Problema pozne integracije so se lotili v [32], kjer so želeli združiti podatke različnih kanalov EEG meritev. Za to so predlagali algoritem, ki uporablja model, ki deluje po principu ansamblov. Ansambel uporablja n osnovnih klasifikatorjev, od katerih je vsak naučen na delu učne množice. Za vsako modalnost se uporabi tak ansambel za gradnjo modela. Končna napoved se izračuna kot uteženo večinsko glasovanje napovedi posameznih modelov (angl. weighted majority voting). Primerjali so točnost napovedi modelov predlaganega algoritma, zgrajenih na podatkih posamezne modalnosti in podatkih, združenih v eno tabelo (zgodnja integracija), z modelom, ki je uporabljal zlivanje vseh modalnosti. Model z zlivanjem je bil značilno boljši od vseh ostalih.

Poglavje 3

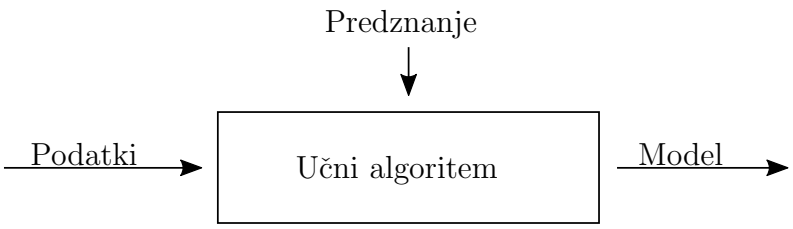
Pregled uporabljenih metod

V raziskavi smo uporabili različne metode nadzorovanega in nenadzorovanega strojnega učenja, ki so podrobneje prestavljene v razdelku 3.1. Med metode nadzorovanega učenja spadajo tiste, s katerimi želimo modelirati ciljno spremenljivko. Glede na tip spremenljivke ločimo dve skupini problemov. Če ciljna spremenljika zavzema omejeno število diskretnih vrednosti, ki jim pravimo razredi, gre za klasifikacijski problem. Spremenljivka je lahko tudi zvezna in zavzame katerokoli numerično vrednost. V tem primeru imamo opravka z regresijskim problemom. Cilj metod, ki rešujejo ta tipa problemov, je najti funkcijo, ki preslika prostor atributov v razred oziroma zvezno vrednost. Učni algoritem mora torej na podanih primerih, ki imajo znano ciljno vrednost, zgraditi model, ki ga lahko nato uporabimo za napovedovanje novih primerov, kot to prikazuje slika 3.1.

Poseben primer nadzorovanega učenja so večciljni modeli, ki so zmožni napovedati več ciljnih spremenljivk naenkrat. To je možno z dvema shemama učenja. Za vsako ciljno vrednost lahko zgradimo svoj enociljni model ali pa uporabimo algoritem, ki že v osnovi zgradi model, s katerim napoveduje vektor ciljnih spremenljivk. Slednji se v praksi pogosto izkaže za boljšega, saj lahko upošteva interakcije med ciljnimi spremenljivkami.

Naloga metod nenadzorovanega učenja je odkriti strukturo v podatkih, ki niso označeni, kar nam pripomore učne primere razdeliti v skupine. To informacijo lahko izluščijo metode gručenja, pri katerih je lahko število želenih skupin podano vnaprej ali pa ga mora algoritem določiti sam.

V izogib prevelikemu prilagajanju algoritmov učnim primerom je potrebno pred kakršnim koli učenjem prostor atributov zmanjšati s ciljem izboljšanja posploševanja, odstranjevanja redundatnih in nekoristnih atributov. Z uporabo metod za filtriranje atributov ocenjujemo kvaliteto posameznega atributa. Delež najbolje ocenjenih nato uporabimo za učenje. Filtrirne metode, uporabljene v naši raziskavi, so opisane v razdelku 3.3.



Slika 3.1: *Algoritem za nadzorovano učenje iz primerov.*

3.1 Algoritmi za nadzorovano učenje

3.1.1 Enociljni modeli za nadzorovano učenje

Logistična regresija

Logistična regresija (angl. logistic regression) je, kljub svojemu zavajajočemu imenu, klasifikator. Ta klasifikator modelira povezavo med atributi x_i, $i = 1, ..., n$ in razredom y, ki lahko zavzame vrednost 0 ali 1. Logistična regresija je v osnovi linearni model, ki uporablja uteženo vstoto atributov za napovedovanje ciljne spremenljivke. Rezultat utežene vsote atributov je zvezna vrednost:

$$\hat{y} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n = \beta^T x, \tag{3.1}$$

ki jo diskretiziramo z logistično funkcijo oziroma izrazimo kot verjetnost, da primer pripada razredu 1:

$$P(y = 1|x, \beta) = \frac{1}{1 + e^{-\hat{y}}} = \frac{1}{1 + e^{-\beta^T x}}. \tag{3.2}$$

Naivni Bayes

Naivni Bayes (angl. naive Bayes) je verjetnostni klasifikator, ki temelji na Bayesovem teoremu. Po Bayesovem pravilu je verjetnost, da primer z vrednostmi atributov x_i, $i = 1, ..., n$ pripada razredu c, enaka:

$$P(c|x_1, x_2, ..., x_n) = \frac{P(c)P(x_1, x_2, ..., x_n|c)}{P(x_1, x_2, ..., x_n)}. \tag{3.3}$$

Naivnost klasifikatorja pride iz predpostavke pogojne neodvisnosti atributov x_i pri danem razredu y_k, kar prejšnji izraz poenostavi v:

$$P(c|x_1, x_2, ..., x_n) = P(c) \prod_{i=1}^{n} \frac{P(c|x_i)}{P(c)}. \tag{3.4}$$

Bayesov klasifikator torej izračuna pogojne verjetnosti za vsak posamezni razred pri danih vrednostih atributov za novi primer, ki ga želimo klasificirati. Primer klasificira v razred z največjo verjetnostjo [33, 34]:

$$y = \underset{c}{argmax} P(c|x_1, x_2, ..., x_n). \tag{3.5}$$

Metoda k-najbližjih sosedov

Ideja metode k-najbližjih sosedov (angl. k-nearest neighbours) je najti vnaprej določeno število učnih primerov k, ki so po razdalji najbližji novemu primeru, in napovedati ciljno vrednost iz njih. Učenje poteka na "len" način, kar pomeni torej samo shranitev učnih primerov oziroma njihove podmnožice.

V primeru klasifikacije v učni množici poiščemo k najbližjih primerov u_i, $i = 1, ..., k$ novemu primeru in napovemo večinski razred. To je razred c, ki mu pripada največ najbližjih primerov:

$$y = \underset{c}{argmax} \sum_{i=1}^{k} \delta(c, c(u_i)), \tag{3.6}$$

kjer je

$$\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{3.7}$$

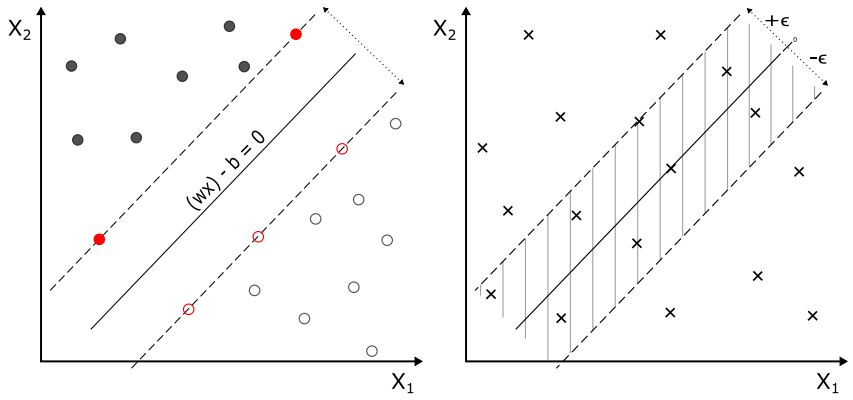
Pri regresiji napovemo kar povprečno vrednost odvisne spremenljivke k najbližjih primerov:

$$y = \frac{1}{k} \sum_{i=1}^{k} y(u_i). \tag{3.8}$$

Metoda je občutljiva na izbrano metriko za računanje razdalj med primeri. Ponavadi se za to uporablja evklidska razdalja [34]. Vseeno lahko najdemo primere, kjer se uporablja kakšna druga metrika, na primer Hammingova razdalja pri klasifikaciji besedil.

Metoda podpornih vektorjev

Metoda podpornih vektorjev (angl. support vector machine) je med najuspešnejšimi metodami za klasifikacijo in regresijo. V osnovi je bila namenjena ločevanju dveh razredov. Metoda poišče takšno hiperravnino, ki v danem prostoru atributov x_i $i = 1, ..., n$ optimalno



(a) *Klasifikacijski SVM* ***(b)*** *Regresijski SVM*

Slika 3.2: *Slika prikazuje ločitveni hiperravnini za obe različici metode podpornih vektorjev.*

ločuje primere dveh razredov. Optimalna hiperravnina je enako in najbolj oddaljena od najbližjih primerov iz obeh razredov. To ravnino lahko podamo z enačbo:

$$(wx) - b = 0. \tag{3.9}$$

Primeri, ki so najbližji hiperravnini, se imenujejo podporni vektorji, razdalja od hiperravnine do podpornih vektorjev pa rob. Cilj je maksimizirati to razdaljo [34]. Slika 3.2a prikazuje hiperravnino s podpornimi vektorji, označenimi z rdečo, ki so na robovih.

Pri regresijski različici je problem zastavljen malo drugače. Metoda mora najti takšno funkcijo $f(x)$, ki se za največ ϵ razlikuje od ciljnih vrednosti y_i učnih primerov. K ceni kriterijske funkcije, ki jo minimiziramo, prispevajo samo primeri zunaj robu [35]. Slika 3.2b prikazuje to razdelitev primerov. Primeri, zunaj območja, označenega s sivimi črtami, prispevajo k računanju cene kriterijske funkcije.

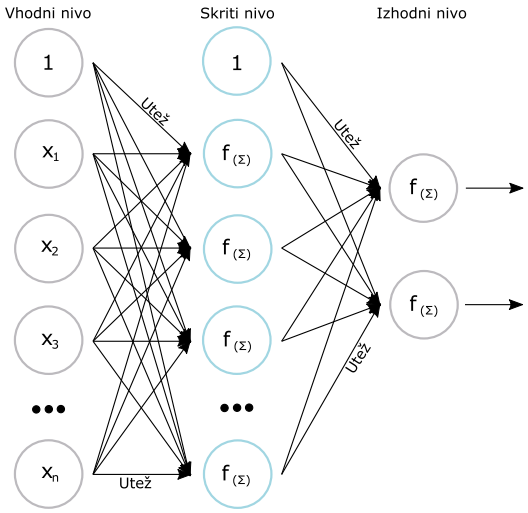
Velika prednost metode podpornih vektorjev je možnost implicitne transformacije danega atributnega prostora v kompleksnejši prostor z uporabo jeder. Tak prostor je lahko bolj primeren za linearno modeliranje, predvsem če so interakcije med atributi v originalnem prostoru nelinearne narave. To lahko naredimo s t.i. jedrnimi funkcijami (radialna funkcija, sigmoidna funkcija, ...).

Umetna nevronska mreža

Umetna nevronska mreža (angl. artificial neural network) je sistem, ki posnema biološki živčni sistem. Zgrajena je iz majhnih neodvisnih gradnikov, imenovanih nevroni, ki so med seboj povezani v mrežo, ki predstavlja končni usmerjeni aciklični graf (glej sliko 3.3).

Mreža je torej sestavljena iz večih zaporedno povezanih nivojev, kar pomeni, da je izhod nevrona na nivoju m lahko le vhod v nevrone na nivoju $m + 1$.

Nevronska mreža je razdeljena na en vhodni nivo, enega ali več skritih nivojev in en izhodni nivo. Število nevronov na vhodnem nivoju je določeno s številom atributov učne množice, topologija vmesnih nivojev je spremenljiva, število nevronov na izhodnem nivoju, pa je odvisno od tipa problema. Z nevronskimi mrežami lahko rešujemo regresijske ali klasifikacijske probleme, med katere spada tudi večrazredna klasifikacija. Pri zadnji je število nevronov na zadnjem nivoju enako številu regresijskih spremenljivk oziroma številu razredov.



Slika 3.3: *Slika prikazuje primer zgradbe nevronske mreže z enim skritim nivojem in dvema izhodnima nevronoma. Vsaka povezava med nevroni ima svojo utež, ki se uporabi za izračun izhoda nevrona po enačbi 3.10.*

Vsak nevron pomnoži vhodne vrednosti s trenutnimi utežmi, prišteje pristranskost in vse skupaj sešteje v eno vrednost, kot je vidno na sliki 3.3. Za vnos nelinearnosti ta pretvori rezultat s pomočjo aktivacijske funkcije $f(x)$ v končni izhod nevrona:

$$X_{out} = f\left(\sum_i w_i x_{in_i} + w_{bias}\right) \tag{3.10}$$

Pogoste aktivacijske funkcije so sigmoidna funckija, Softmax, hiperbolični tangens in ReLU (angl. Rectified Linear Unit). Na izhodnem nivoju je možna tudi uporaba identitete kot aktivacijske funkcije. Tako dobimo regresijsko nevronsko mrežo.

V fazi učenja se mora nevronska mreža naučiti uteži w_i iz enačbe (3.10) za vsak nivo. Cilj je čimbolj natančna preslikava iz atributnega prostora v prostor ciljne spremenljivke.

Za to se uporabljajo algoritmi optimizacije. Najpogosteje uporabljen je algoritem z vzratnim razširjanjem napake.

Odločitvena drevesa

Osnovna naloga odločitvenega drevesa (angl. decision tree) je razdelitev prostora atributov na več manjših podprostorov. To razdelitev rekurzivno ponavljamo, dokler niso podprostori tako majhni, da lahko na njih zgradimo preproste modele. Ponavadi je to konstantana vrednost, lahko pa jo nadomestimo s preprostim modelom, kot je linearni. Zgrajeno odločitveno drevo sestavljajo notranja vozlišča, ki ustrezajo atributom, veje, ki ustrezajo podmnožicam vrednosti atributov, in listi, ki predstavljajo majhne podprostore v celotnem prostoru atributov.

Vsako nekončno vozlišče na posameznem nivoju predstavlja najbolj informativen atribut, ki ga izberemo s pomočjo ustreznega kriterija. Pri klasifikacij je to naprimer Gini indeks, pri regresiji pa povprečni kvadrat napake.

Pri napovedovanju ciljne vrednosti novega primera se na podlagi vrednosti atributov sprehodimo od korena navzdol do lista, ki določa funkcijo za ustrezno napoved. Ta je ponavadi kar povprečje vrednosti primerov pri regresiji oziroma večinski razred pri klasifikaciji.

Največji prednosti odločitvenih dreves sta enostavno razumevanje in interpretacija, saj se jih da med drugim tudi vizualizirati. Med drugim je gradnja razmeroma hitra in enostavna. Gradnja ni omejena na noben tip podatkov.

Po drugi strani lahko algoritem za gradnjo drevesa zgradi preveč kompleksno drevo, ki ne generalizira podatkov dovolj dobro. Poleg tega je drevo zelo občutljivo na majhne spremembe v učni množici, ki lahko posledično spremenijo strukturo drevesa.

Naključni gozdovi

Naključni gozdovi (angl. random forest) deljujejo na principu ansamblov. Ansambli zgradijo množico preprostih modelov s spreminanjem notranjih parametrov ali z izbiro različnih podmnožic učnih podatkov. Različni modeli tako predstavijo podatke iz različnih pogledov. Končna napoved je združena napoved posameznih modelov.

Naključni gozd sestavlja množica odločitvenih dreves. Vsakega od teh modelov dobimo tako, da vpeljemo stohatičnost v postopek gradnje. Za posamezen model iz celotne učne množice izberemo k primerov kot permutacije učnih primerov s ponavljanjem. Ta nova podmnožica se uporabi za gradnjo drevesa, tako da se v vsakem vozlišču izbere podmnožica atributov, na podlagi katerih se izračuna najbolj informativnega in tako rekurzivno razdeli vozlišče.

Pri klasifikaciji novega primera gozd zbere glasove posameznih dreves in napove razred, kamor primer uvršča večina klasifikacijskih dreves. Pri regresiji se napovedi posameznih regresijskih dreves povprečijo. Takšno združevanje napovedi zmanjša varianco posameznih odločitvenih dreves in robustno napove končno vrednost.

3.1.2 Večciljni modeli za nadzorovano učenje

Umetna nevronska mreža

Umetna nevronska mreža ima lahko po definiciji več izhodnih nevronov. To lahko, tako kot pri večrazredni klasifikaciji, izkoristimo tudi pri večciljnem napovedovanju. Za vsako ciljno vrednost lahko predvidimo namenski izhodni nevron. Pri tej razširitvi si ciljne spremenljivke delijo notranje povezave nevronske mreže. S tem pride do deljenja informacij med ciljnimi spremenljivkami, kar lahko izboljša napovedno moč modela, saj se lahko deli mreže specializirajo za napovedovanje posamezne ciljne spremenljivke, obenem pa izluščijo skupne informacije.

Za učenje vseh nevronskih mrež je v splošnem potrebno normalizirati atribute, tako da zavzemajo vrednosti med 0 in 1. To storimo zato, da je učenje bolj stabilno. Izhod iz nevronske mreže prav tako zavzema vrednosti na tem intervalu. Za večciljno nevronsko mrežo moramo v času učenja tudi ciljne spremenljivke normalizirati. Pri napovedovanju novega primera uporabimo parametre, uporabljene pri normalizaciji posamezne spremenljivke in napovedi nevronske mreže umestimo na originalne intervale.

Odločitvena drevesa

Odločitvena drevesa je možno posplošiti za napovedovanje več ciljnih spremenljivk. Listi v tem primeru vsebujejo funkcijo za napoved več vrednosti. Primer take funkcije je povprečje vrednosti primerov pri regresiji oziroma večinski razred pri klasifikaciji za posamezno ciljno spremenljivko. Pri napovedovanju ciljne vrednosti novega primera veljajo enaka pravila kot pri enociljnih odločitvenih drevesih. Edina razlika je, da je napoved vektor, ki vsebuje napovedi za vsako ciljno spremenljivko posebej.

Kriterij za delitev primerov v nekončnih vozliščih pri večciljnem drevesu upošteva vse ciljne spremenljivke. Za izbiro najbolj informativnega atributa uporabimo povprečje vrednosti izbrane mere preko vseh posameznih ciljnih spremenljivk. Taka rešitev ni primerna pri mešanju klasifikacijskih in regresijskih nalog, saj se mere za ocenjevanje atributov niso primerljive med seboj. Tam je potrebna uporaba mere, ki za vsako ciljno spremenljivko rangira atribute glede na pripadajočo mero [36].

Naključni gozdovi

Naključni gozd pri večciljnih problemih deluje podobno kot pri enociljnih problemih. Sestavljajo ga večciljna odločitvena drevesa, zato lahko tudi gozd napovedove več hkratnih ciljnih spremenljivk. Pri novi napovedi se za vsako ciljno spremeljivko uporabi najpogosteje napovedana vrednost preko vseh dreves (klasifikacija) ali pa se izračuna povprečna vrednost napovedi posameznih dreves (regresija).

3.2 Algoritmi za nenadzorovano učenje

Metoda k-voditeljev

Metoda k voditeljev (angl. k-means) je eden izmed najpreprostejših algoritmov, ki rešujejo problem gručenja. Metoda poskuša razdeliti učne primere x_i, $i = 1...n$ v k različnih skupin, tako da je razdalja od posameznega primera do centra skupine najmanjša možna. Poskuša torej najti take centre skupin, ki minimizirajo varianco znotraj posamezne skupine. Število skupin je določeno vnaprej.

Iskanje točne rešitve za poljubne podatke predstavlja NP-težek problem, zato namesto tega iščemo približno rešitev. Seveda to pomeni, da lahko algoritem vrne rešitev, ki je neprimerna. Zato se ponavadi algoritem izvaja večkrat z različnimi začetnimi postavitvami centroidov skupin. Od začetne lokacije centroidov sta odvisna čas izvajanja algoritma in kvaliteta končne rešitve. To se lahko naredi naključno ali pa s kakšnim bolj inteligentnim algoritmom, kot je KMeans++, ki poskrbi, da so začetni voditelji bolj oddaljeni drug od drugega in postavljeni bližje centrom skupin.

3.3 Metode za izbiro atributov

ReliefF in RReliefF

Algoritmi iz družine Relief veljajo za najuspešnejše za izračun kakovosti atributov. Kakovost atributov ocenijo glede na to, kako dobro njihove vrednosti ločijo med primeri, ki so si blizu. Metode znajo izrabiti lokalne informacije o primerih in pravilno oceniti kvaliteto med seboj močno odvisnih atributov [37].

Originalna metoda Relief je omejena na dvorazredno klasifikacijo, ni zmožna obravnavati manjkajočih podatkov in ni robustna za uporabo s šumnimi podatki. Nadgradnja ReliefF te pomanjkljivosti odpravi. Za izračun kakovosti atributov ReliefF n-krat naključno izbere primer iz podatkovne zbirke in za vsak razred posebej poišče njemu k najbližjih primerov. Za izračun prispevka vsakega najbližjega primera se uporabi enačba, ki upošteva

tip atributa (diskretni ali zvezni) in morebitne manjkajoče vrednosti. Povprečja vrednosti prispevkov posameznih atributov se uporabijo za izračun končne ocene posameznega atributa.

Za regresijske probleme uporabljamo nadgradnjo – algoritem RReliefF. Pri regresijski problemih seveda ni mogoče določiti pripadnosti posameznega primera razredu. Namesto tega RReliefF uporablja verjetnost, da sta spremenljivki dveh primerov različni, ki jo modelira z relativno razdaljo med vrednostima. Za izračun kakovosti RReliefF, tako kot ReliefF, n-krat naključno izbere primer iz podatkovne zbirke. Za razliko od ReliefFa, RReliefF poišče samo k najbližjih primerov in na njih modelira verjetnosti, da so vrednosti spremenljivk različne. Končna ocena posameznega atributa se izračuna z enačbo, ki upošteva verjetnost, da so vrednosti atributa različne, verjetnost, da so ciljne vrednosti različne, in verjetnost, da so tako vrednosti atributa kot vrednosti ciljnih spremenljivk različne [37, 38].

Laplaceova ocena

Laplaceova ocena se uporablja za ocenjevanje atributov v zbirkah podatkov, ki nimajo znanih ciljnih spremenljivk. Metoda bazira na ideji, da dva primera pripadata istemu razredu oziroma skupini, če sta si v prostoru atributov blizu. Lokalna struktura prostora atributov je torej pomembnejša od globalne, kot je to značilno za veliko algoritmov strojnega učenja. Lokalno geometrijsko strukturo metoda modelira z grafom najbližjih sosedov, s katerim oceni kvaliteto atributov. Atribute, zaradi katerih sta dva primera blizu v grafu, metoda oceni kot kvalitetne [39].

Laplaceova ocena za r-ti atribut se izračuna po enačbi:

$$L_r = \frac{\sum_{i,j}(f_{ri} - f_{rj})^2 S_{ij}}{\sum_i (f_{ri} - \mu_r)^2 D_{ii}}, \tag{3.11}$$

kjer je μ_r povprečje primerov r-tega atributa, D diagonalna matrika z $D_{ii} = \sum_j S_{ij}$ in S je matrika podobnosti, kjer so elementi definirani po enačbi:

$$S_{ij} = \begin{cases} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}, & \text{če sta } x_i \text{ in } x_j \text{ soseda} \\ 0, & \text{drugače.} \end{cases} \tag{3.12}$$

Pri tem je σ konstanta, x_i pa posamezen primer v podatkovni množici.

Iz enačbe (3.11) je razvidno, da z minimizacijo imenovalca želimo dodeliti manjšo oceno tistim atributom, ki spoštujejo vnaprej definirano strukturo grafa. Torej, večji kot je S_{ij}, bližje sta si primera v grafu, in manjša kot je razlika med posameznima vrednostima f_{ri} in f_{rj}, bolj sta si primera podobna v tem atributu. Atributi, ki to spoštujejo preko vseh primerov, so ocenjeni z dobro oceno.

Poglavje 4

Zlivanje podatkovnih množic

Za skoraj vsak problem lahko zberemo podatke, zajete z različnimi meritvami (modalnostmi). Podatki so lahko tudi direktno povezani s problemom ali pa opisujejo nek drug problem, ki je posredno soroden našemu. Demenco, na primer, lahko povzročijo različne bolezni in poškodbe. Naša naloga je ugotoviti, ali je Alzheimerjeva bolezen razlog za njen nastanek. Za to potrebujemo različne podatke o pacientu: rezultate kognitivnih testov, rezultate laboratorijskih testov, medicinske slike, pretekle bolezni in poškodbe, demografske podatke, podatke o življenskem okolju, podatke o družinskih boleznih, navadah itd. Taka množica podatkov nam pomaga odkriti skrite vzorce in izločiti vzroke, ki bi bili lahko razlog za nastanek bolezni. Če je oseba doživela hudo poškodbo glave ali pa ima mogoče Parkinsonovo bolezen, je verjetnost, da je vzrok demence Alzheimerjeva bolezen, precej manjša. Prav tako bi lahko, na primer, stresno delovno okolje vplivalo na nastanek bolezni.

Za integracijo informacij vseh različnih virov podatkov, ki so na voljo, se uporablja zlivanje. Glede na [40, 41] lahko algoritme za zlivanje razdelimo v tri skupine, glede na to, v katerem delu učenja napovednega modela se izvede. Pri zgodnji integraciji se pred gradnjo napovednega modela zgradi združeno podatkovno množico, ki jo nato uporabimo kot vhod v algoritem za gradnjo modela. Po navadi gre za preprosto konkatenacijo podatkov po nekem ključu. Če tak ključ ne obstaja, je potrebno uporabiti algoritem, s katerim lahko dopolnimo mankajoče atribute. Pri vmesni integraciji so posamezne podatkovne množice vhod v algoritem strojnega učenja. Pri tem algoritem ne združi podatkovnih množic v eno skupno, ampak iz vsakega posameznega izlušči informacije, ki jih uporabi za gradnjo modela. Pri pozni integraciji za vsako podatkovno množico zgradimo svoj napovedni model. Napovedi posameznih modelov nato združimon.

V tej nalogi smo raziskovali vpliv zlivanja na klasifikacijsko točnost pri učenju iz različnih podatkovnih množic. Glede na pregled področja je apliciranje teh metod za

raziskovanje Alzheimerjeve bolezni še v povojih. Naš pristop sledi principu zgodnje integracije in je opisan v razdelku 4.1. Metoda temelji na hipotezi, da lahko z dodatnimi atributi, pridobljenimi z zlivanjem iz druge množice, pridobimo koristne informacije, s katerimi lahko izboljšamo napovedni model in posledično zvišamo napovedno točnost.

4.1 Zlivanje podatkov na osnovi skupnih atributov

Glavna ideja naše metode je združiti podatkovne množice preko enakih atributov. Slika 4.1 prikazuje postopek zlitja dveh podatkovnih množic. V prikazanem primeru množici A priredimo atribute iz množice B in tako dobimo razširjeno množico, na kateri lahko naučimo model za napovedovanje bolezni. Metoda je sestavljena iz 5 korakov:

1. Metoda najprej poišče presek atributov podatkovnih množic A in B. Tako lahko razdelimo množico A na dva dela X in Y_1 in množico B na Z in Y_2.

2. Podmnožico Z ocenimo z algoritmom ReliefF oziroma Laplaceovo oceno, da s tem izberemo k najboljših za zlitje v množico A. Izbira algoritma za oceno atributov je odvisna od prisotnosti ciljne spremenljivke v množici B.

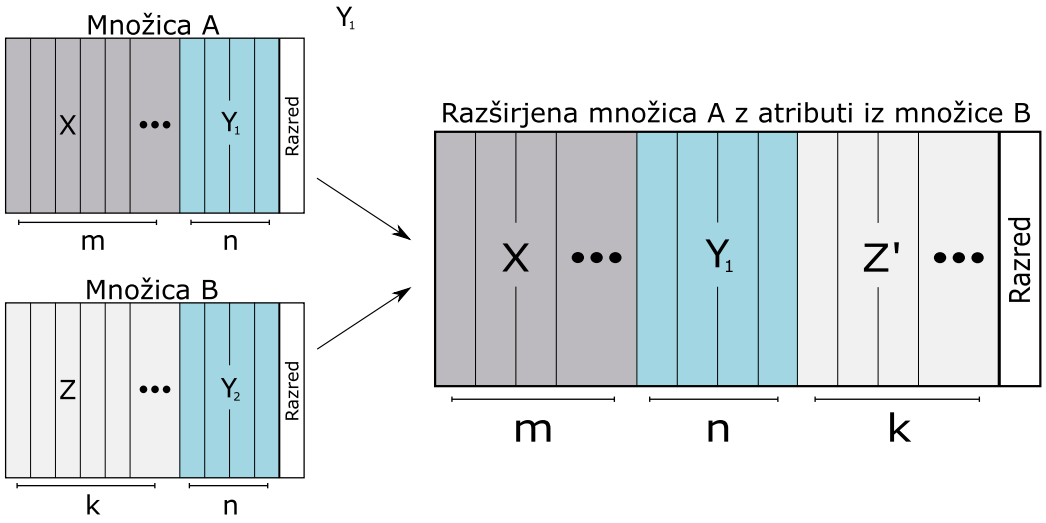
3. Skupno podmnožico atributov Y_2 najprej ocenimo in nato izberemo n najboljših z uporabo algoritma RReliefF.

4. Na selektivno izbranih atributih podmnožice Y_2 nato gradimo enega ali več napovednih modelov, ki modelirajo odvisnost med atributi Y_2 in vsakim posameznim atributom iz množice Z.

5. Z uporabo atributov Y_1 in zgrajenih modelov nato napovemo vrednosti manjkajočih atributov v množici Z'.

Rezultat je zlitje množice B v množico A, ki ga predstavlja podatkovna množica $[X|Y|Z']$. Možna je seveda tudi obrnjena situacija - torej zlitje množice A v množico B. V tem primeru dobimo zlito množico $[Z|Y|X']$.

Zlita množica je tako pripravljena na nadaljnjo analizo in napovedovanje Alzheimerjeve bolezni. Na tej množici je prav tako zaradi prevelike dimenzionalnosti potrebna izbira atributov, ki jo naredimo z algoritmom ReliefF. Na izhodni množici nato zgradimo različne osnovne modele, opisane v razdelku 3.1.



Slika 4.1: *Zlitje dveh podatkovnih množic preko skupnih atributov.*

4.2 Modeli za napovedovanje manjkajočih atributov

V fazi zlivanja smo uporabili dve različni vrsti napovednih modelov. Narava problema je takšna, da moramo napovedati več ciljnih spremenljivk z isto množico atributov. To lahko naredimo na dva načina. Za vsako ciljno spremenljivko lahko zgradimo svoj model ali pa uporabimo model, ki je zmožen napovedati več ciljnih spremenljivk hkrati. Oba pristopa imata svoje prednosti in slabosti. Enociljni modeli se pri učenju osredotočijo na eno samo spremeljivko in poskušajo zgraditi model, ki se bo kar najbolje naučil ene odvisnosti. Tako lahko dosežemo, da je napovedovanje posamezne spremenljivke čimbolj neodvisno. To je pomembno predvsem, ko si ciljne spremenljivke niso sorodne. Nasprotje temu je uporaba večciljnega modela, ki je kompromis preslikav atributov v vse ciljne spremenljike hkrati. Po eni strani je tak model bolj splošen, saj lahko izlušči informacijo o medsebojni odvisnosi ciljnih spremenljivk. Po drugi strani pa lahko prekomerno število ciljnih spremenljivk, ki si niso sorodne, in premajhno število primerov poslabšata natačnost takega modela.

Enociljni model v našem primeru modelira preslikavo atributov iz množice Y_2 po enačbi:

$$f\left(\boldsymbol{Y_2}\right) = z \in Z, \tag{4.1}$$

kjer je z en atribut iz množice Z. Za napovedovanje vrednosti vseh atributov iz Z moramo torej zgraditi k modelov, kjer je k število atributov iz množice Z.

V primeru uporabe večciljnega modela preslikamo vrednosti iz prostora atributov množice Y_2 v prostor vseh atributov množice Z. Rezultat preslikave je torej vektor spremenljivk, ki ga modelira enačba:

$$f\left(\boldsymbol{Y_2}\right) = \boldsymbol{Z}, \tag{4.2}$$

kjer je $\boldsymbol{Z}$ vektor vseh atributov množice Z.

Ciljne spremeljivke so lahko v splošnem diskretne ali zvezne. Od tega je odvisno, ali za modeliranje odvisnosti med atributi in ciljno spremenljivko uporabimo klasifikacijski ali regresijski model. V našem primeru smo to poenostavili, tako da smo diskretne ciljne spremenljivke binarizirali v zvezne spremenljivke in jih tako pripravili za uporabo v regresijskih modelih.

Uporabili smo obe metodologiji za grajenje modelov, ki znata napovedati vektor ciljnih spremenljivk. Pri metodologiji, ki zgradi za vsako ciljno spremenljivko svoj enociljni regresijski model, smo uporabili tri različne algoritme. Dva sta bila algoritma nazdorovanega učenja, eden pa je temeljil na algoritmu nenadzorovanega učenja. V prvo skupino sta spadala naključni gozdovi in metoda podpornih vektorjev, v drugo pa algoritem, ki je temeljil na metodi k voditeljev. Algoritem je na množici Y_2 najprej poiskal gruče in vsak primer iz te množice dodelili eni izmed njih. Na podlagi tega smo za vsako gručo izračunali povprečje ciljnih vrednosti pripadajočih primerov. V času napovedi smo to povprečje uporabili kot rezultat preslikave.

Druga metodologija je izrabljala moč večciljnih regresijskih modelov. Za dopolnjevanje atributov smo uporabili prilagojene algoritme za umetne nevronske mreže, naključni gozd in odločitveno drevo.

Poglavje 5

Eksperimentalno okolje

V nalogi smo ovrednotili našo metodo za fuzijo na več različno velikih množicah. Na združenih množicah smo naučili večje število osnovnih klasifikatorjev in primerjali rezultate z osnovnimi množicami, da bi ugotovili, ali atributi, pridobljeni z zlivanjem, prispevajo k zvišanju klasifikacijske točnosti. V naslednjih poglavjih opišemo uporabljene podatkovne množice, postopek učenja in način vrednotenja napovedi.

5.1 Postopek učenja

Vse klasifikatorje smo testirali z 10-kratnim prečnim preverjanjem. Pri razdelitvi osnovne množice na učni in testni del smo poskrbeli za enako porazdelitev vrednosti ciljne spremenljivke. Prav tako smo poskrbeli, da so imeli vsi klasifikatorji enako razdelitev množice na učni in testni del v vsakem od preverjanj (kar je potrebno pri statistični evalvaciji). Pred učenjem modela smo naredili tudi izbor atributov.

Za zlivanje množic smo znotraj naše metode uporabili različne učne algoritme. Za izbiro najboljšega smo uporabili postopek 10-kratnega notranjega prečnega preverjanja. Znotraj vsake razdelitve osnovne množice na učni in testni del smo učno množico razdelili na nov učni in validacijski del. Na ta način smo izbrali optimalni algoritem za nadzorovano učenje in ga uporabili v algoritmu za zlivanje množic na celotni učni množici.

Za vse algoritme strojnega učenja smo uporabili privzete nastavitve, analizo njihovega vpliva pa smo prepustili nadaljnjemu delu. Prav tako smo na začetku omejili število atributov, ki smo jih dodali osnovni množici, na 50, in število atributov za učenje napovednega modela na 100. Obe števili sta bili izbrani bolj kot ne naključno, zato da bi zmanjšali čas izvajanja algoritmov in seveda prekomerno prilagoditev podatkov učni množici.

5.2 Uporabljene podatkovne množice

V literaturi se pojavljajo podatki iz različnih virov. Velikokrat gre za podatke kakšnih manjših študij, ki niso javno dostopni. Takšne podatkovne zbirke so posledično precej okrnjene, saj vsebujejo podatke za le okoli 100 različnih oseb ali manj. Tipi podatkov so v takih primerih večinoma omejeni na demografske podatke in rezultate kognitivnih testov.

V preostali literaturi lahko zasledimo uporabo podatkov iz večjih študij. Podatkovne zbirke teh študij so bolj raznolike in vsebujejo podatke različnih modalnosti. Take študije so večinoma longitudinalne in v daljšem obdobju pridobijo večje število oseb.

Pogosti tipi podatkov, ki se pojavljajo, so klinični podatki. Ti vsebujejo osebne informacije o osebah. To vključuje demografske podatke (spol, starost, stopnjo izobrazbe ipd.), družinsko zgodovino (družinske bolezni, podatke o sorodnikih ipd.) in rezultate testov, ki preverjajo motorične, senzorske in kognitivne sposobnosti. Ponavadi zdravniki izvedejo več testov za čim natanjčnejšo diagnozo. Najpogostejši testi so MMSE (angl. Mini Mental State exam), NPI (angl. Neuropsychiatric Inventory), CDR (angl. Clinical Dementia Rating), Camdex (angl. Cambridge Mental Disorders of the Elderly Examination), DEMQOL (angl. Health related quality of life test), test za preverjanje motenj motoričnih in zaznavnih sposobnosti itd. Nekatere od teh izvedejo tudi s skrbniki, da ugotovijo, kako bolezen vpliva nanje. To so predvsem testi o kvaliteti življenja.

Biomarkerji iz krvi in cerebrospinalne tekočine vsebujejo meritve količine proteinov. Najpogosteje sta izmerjena proteina tau in amiloid beta, ker njuno nabiranje v možganih potrjuje Alzheimerjevo bolezen.

Medicinske slike so v zbirkah redko obdelane. Obstajajo trije pristopi za pridobitev atributov iz slik. Pri prvem se kot atributi direktno uporabijo intenzitete vokslov. Drugi pristop iz vokslov izračuna različne atribute, kot so debelina skorje, globina režnjev itd. Pri tretjem se upošteva logične skupine vokslov, za katere se izračuna atribute, kot je, npr. volumen.

V naši raziskavi smo uporabili podatke iz treh različnih virov, ki so opisani v razdelku 5.2.1. V razdelku 5.2.2 predstavimo programe, s katerimi smo obdelali slike MRI in iz njih izluščili atribute.

5.2.1 Uporabljeni viri podatkov

Podatkovna zbirka AddNeuroMed

Študija AddNeuroMed je rezultat javno-zasebnega partnerstva za odkrivanje biomarkerjev in replikacijo Alzheimerjeve bolezni. Študijo so izvajali v 6 različnih evropskih institucijah iz Italije, Finske, Grčije, Velike Britanije, Poljske in Francije. V obdobju od januarja 2006 do februarja 2008 je bilo v študijo vključenih 781 oseb. To je bila longitudinalna študija,

kjer so vsako osebo spremljali daljše obdobje. Po prvem pregledu je imela vsaka oseba preglede na 3 mesece. Vsaka oseba je v povprečju opravila le nekaj preiskav, nobena se ni udeležila vseh [42].

Podatkovno zbirko sestavljajo slike MRI možganov, analize genov, analize krvnih proteinov, rezultati kliničnih in kognitivnih testov. Kljub temu, da je bilo v študijo vključenih 781 oseb, je za modalnosti, kot sta slike MRI in analize krvnih proteinov, na voljo le omejena količina podatkov. Slike MRI so na primer na voljo le za 163 oseb iz prvega obiska, medtem ko so analize krvnih proteinov na voljo za 659 oseb iz prvega obiska in 272 oseb iz kasnejših obiskov [42].

Ob vsakem obisku so osebam postavili diagnozo glede na rezultate kognitivnih testov. Diagnozo sestavljajo 3 kategorije:

- Normalno stanje (266 oseb)
- Blaga kognitivna motnja (257 oseb)
- Alzheimerjeva bolezen (258 oseb)

V kategorijo *normalno stanje* so bile uvrščene osebe brez nevroloških ali psihiatričnih motenj. V kategorijo *blaga kognitivna motnja* so bile uvrščene osebe, ki so izkazovale probleme s kognitivnimi sposobnostmi, vendar jih to ni oviralo pri vsakdanjih aktivnostih. V kategorijo *Alzheimerjeva bolezen* so bile uvrščene vse osebe s kritičnimi kognitivnimi motnjami v vsaj dveh različnih področjih, kot so spomin, orientacija, reševanje problemov itd. Pri tem ni bilo prisotne nobene druge bolezni, ki bi lahko povzročila demenco [43, 44].

Srbska podatkovna zbirka

Raziskovalci iz Univerze v Novem Sadu so zbrali tri podatkovne množice. Ena je direktno povezana z Alzheimerjevo boleznijo, preostalo dve pa z demenco in možgansko kapjo. Vse tri množice vsebujejo demografske podatke o osebah. Poleg tega vsebujejo različne teste kognitivnih, motoričnih in senzoričnih sposobnosti in teste, ki preverjajo sposobnost samooskrbe. Množica s podatki o možganski kapi vsebuje tudi podatke o poškodbah možganov za različne možganske regije. Vse skupaj je bilo pregledanih 245 oseb, od tega jih je imelo 29 postavljenih diagnozo Alzheimerjeve bolezni, 27 jih je imelo blažjo kognitivno motnjo in 29 je bilo zdravih oseb. Osebe v množicah o demenci in možganski kapi niso imele postavljene diagnoze o prisotnosti/odsotnosti Alzheimerjeve bolezni.

Podatkovna zbirka OASIS

OASIS (angl. Open Access Series of Imaging Studies) je projekt, s katerim želijo omogočiti raziskovalcem prost dostop do podatkovnih zbirk s slikami MRI. Zbirko so zbrali in izdali

raziskovalci iz Univerze v Washingtonu, Harvardske Univerze, Medicinskega inštituta Howard Hughes in bolnišnice v Massachusettsu.

Podatkovna zbirka vsebuje MR slike 416 oseb starih od 18 do 96 let. Vsako osebo so slikali večkrat in tako zmanjšali razmerje med signalom in šumom. Za razliko od drugih podatkovnih zbirk ta vsebuje slike oseb iz različnih življenjskih obdobij, kar omogoča raziskovanje staranja in ne samo Alzheimerjeve bolezni. Poleg slik MRI so dostopni tudi demografski podatki in rezultati dveh kognitivnih testov [45]. Nobena od oseb v tej množici ni imela postavljene diagnoze o prisotnosti/odsotnosti Alzheimerjeve bolezni.

5.2.2 Predprocesiranje magnetno-resonančnih slik

Surove MR slike je težko uporabiti kot vhodne podatke za algoritme strojnega učenja. Poleg možganov je na slikah tudi lobanja, ki v bistvu predstavlja šum. Zato je obvezno predprocesiranje teh slik. Tudi v literaturi so redki primeri, kjer lahko zasledimo uporabo surovih slik.

V ta namen se pogosto uporabljajo odprtokodni programi FreeSurfer, ANTs in Mindboggle. ANTs in FreeSurfer se uporabljata za procesiranje surovih slik. Med drugim poskrbita za odstranjevanje lobanje, segmentacijo slike v različne vrste tkiv (sivo in belo možganovino in cerebrospinalno tekočino) in označevanje regij možganske skorje in struktur pod možgansko skorjo. FreeSurfer dodatno skonstruira tudi 3D rekonstrukcijo možganske skorje obeh hemisfer. Rezultat procesiranja so tudi različne meritve za vsako od regij. Med njih spadajo debelina možganske skorje in volumetrične meritve, kot sta lokalna ukrivljenost in površina [7, 8, 9].

Razlika med programoma je, da za nekatere od korakov procesiranja uporabljata različne algoritme. ANTs, na primer, za ekstrakcijo možganov in segmentacijo tkiv uporablja algoritme, ki uporabljajo predloge že obdelanih možganov kot dodatno informacijo. FreeSurfer tega ne uporablja, vendar direktno iz segmentirane bele in sive možganovine preračuna željeno topologijo rekonstrukcije. Zaradi tega se lahko končne meritve razlikujejo [9].

Mindboggle na podlagi vmesnih in končnih rezultatov programov ANTs in FreeSurfer izračuna atribute za vsako izmed regij možganov. Za to uporabi segmentacije, izračunane s programoma FreeSurfer in ANTs, in rekonstruirano 3D možgansko skorjo z označenimi regijami. Končni rezultat so atributi, med katere spadajo [8, 46]:

1. volumni vseh označenih regij,
2. debelina možganske skorje označenih regij,
3. statistične povzetke meritev oblik (za vsako regijo):

- površina,
- povprečna ukrivljenost,
- geodezična globina,
- globina sprehoda,
- konveksnost,
- debelina.

V tej raziskavi smo uporabili slike MRI iz podatkovnih zbirk AddNeuroMed in OASIS. Slike AddNeuroMed zbirke so bile že spredprocesirane, rezultati pa dostopni na spletu. Za naše potrebe smo z uporabo zgoraj omenjenih orodij obdelali 167 slik podatkovne zbirke OASIS.

5.2.3 Priprava podatkovnih množic za analizo

Analizirali smo podatke treh različnih virov, opisanih v razdelku 5.2.1. Primeri v posameznih množicah podatkov niso vsebovali podatkov za vse modalnosti. V AddNeuroMed množici so, na primer, manjkali podatki o količini proteinov v krvi za nekatere primere. Takšno množico smo razdelili na več delov, tako da je vsak del vseboval samo modalnosti, za katere so bili podatki dostopni za vse primere. Atribute, ki so po tej razdelitvi še vedno vsebovali neznane vrednosti, smo odstranili. S tem postopkom smo izluščili 8 različnih množic, primernih za poskuse, ki jih lahko vidimo v tabeli 5.1. V tej tabeli so vidni tudi osnovni podatki o množicah.

Pred učenjem smo pregledali vse atribute vseh množic in ugotovili, kateri so diskretni. Te smo nato binarizirali in jih pripravili za uporabo v algoritmih strojnega učenja.

Za določitev parov množic za zlivanje smo najprej poenotili poimenovanje atributov iz vsakega vira in ugotovili možne preseke med množicami. Glede na moč preseka smo določili pare za zlivanje, prikazane v tabeli 5.2. Pri tem smo atribute iz druge množice dodali atributom iz prve.

5.3 Primerjava uspešnosti metod

Uspešnost napovedi posameznih zgrajenih modelov smo merili s klasifikacijsko točnostjo. Uspešnost modelov, naučenih na zlitih in osnovnih množicah smo primerjali medseboj. Da bi ugotovili, ali obstaja značilna razlika med njimi, smo primerjave naredili s statističnimi testi. Ker imamo opravka s primerjavo več klasifikatorjev, smo uporabili neparametrični Friedmanov test v kombinaciji s post-hoc Nemenyi testom. Vsako posamezno metodo

Množica	Razdelitev primerov (NK/AB)	Število atributov	Ciljna spremenljivka
TestAndDemo	833/802	19	✓
TestAndDemo+Proteins	201/272	1035	✓
TestAndDemo+BrainData	117/45	2170	✓
Proteins	274/184	1016	✓
Oasis	Ni podatka	3220	✗
SrbAlzheimer	24/42	135	✓
SrbDementia	Ni podatka	107	✗
SrbStroke	Ni podatka	52	✗

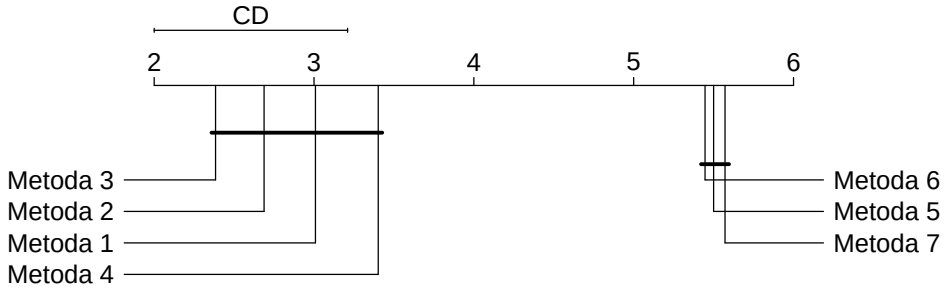
Tabela 5.1: *Tabela prikazuje osnovne podatke o množicah: število atributov, prisotnost ciljne spremenjlivke in število primerov, ki pripadajo Alzheimerjevi bolezni (AB) in normalnim kontrolam (NK).*

Okrajšava	Par		
TTP	TestAndDemo	-	TestAndDemo+Proteins
TTB	TestAndDemo	-	TestAndDemo+BrainData
TBO	TestAndDemo+BrainData	-	Oasis
TBTP	TestAndDemo+BrainData	-	TestAndDemo+Proteins
TPTB	TestAndDemo+Proteins	-	TestAndDemo+BrainData
PTP	Proteins	-	TestAndDemo+Proteins
SASD	SrbAlzheimer	-	SrbDementia
SASS	SrbAlzheimer	-	SrbStroke

Tabela 5.2: *Tabela prikazuje pare za zlivanje. Levi množici smo dodali atribute iz desne množice. Zaradi dolgih imen množic smo za namene lažje predstavitve rezultatov dodali tudi okrajšave.*

zlivanja smo statistično primerjali tudi z metodo brez zlivanja. Za to primerjavo smo uporabili neparametrični Wilcoxonov test.

Za grafično predstavitev rezultatov smo uporabili graf rangov s kritično razdaljo [47]. Na sliki 5.1 je simbolični primer takega grafa, ki primerja fiktivne rezultate 7 metod. Metode so rangirane od leve proti desni, pri čemer je nauspešnejša metoda prva iz leve. Metode, povezane s temno vodoravno črto niso značilno različne. Dolžina črte, označene s CD, ponazarja kritično razdaljo, ki definira, za koliko se morata dve metodi razlikovati, da ju lahko označimo kot statistično značilno različni. Vse teste smo izvedli pri stopnji zaupanja $\alpha = 0.05$.



Slika 5.1: *Primer grafa rangov s kritično razdaljo.*

Poglavje 6

Rezultati in evalvacija

V tem poglavju predstavimo rezultate, jih primerjamo in evalviramo. Analiziramo, ali ima predlagana metoda zlivanja podatkov pozitivni učinek na klasifikacijsko točnost. Klasifikacijske točnosti smo izmerili na osnovni in na zlitih množicah in jih primerjali. Dodatno analiziramo tudi uspešnost posameznih klasifikatorjev, ki smo jih uporabili za gradnjo modelov za napovedovanje Alzheimerjeve bolezni.

6.1 Uspešnost posameznih metod zlivanja

Naša naloga je bila odkriti, ali predlagane metode fuzije izboljšajo napovedno točnost zgrajenega modela. Za gradnjo napovednih modelov smo uporabili linearno regresijo (LR), naivnega Bayesa (NB), odločitveno drevo (DT), k-najbližjih sosedov (KNN), naključni gozd (RF), metodo podpornih vektorjev (SVM) in nevronsko mrežo (NN). Vsakega od napovednih modelov smo zgradili na paru zlitih množic iz tabele 5.2 in samo na osnovni množici (levi član para v tabeli). Za večjo preglednost smo v tabelah v tem poglavju uporabili indikatorje oblike par množic-klasifikator.

Tabela 6.1 prikazuje klasifikacijske točnosti za vse možne kombinacije trojic **par množic - klasifikator - metoda zlivanja**. Iz nje je razvidno, da je uporaba zlivanja v nekaterih primerih izboljšala napovedno točnost, v nekaterih pa celo poslabšala. Na primer, pri SASD-NB in SASS-NB se je napoved po zlivanju (večciljno in enociljno regresijsko zlivanje) izboljšala tudi do 11%, medtem ko se je v primeru SASS-SVM po zlivanju poslabšala (enociljno zlivanje z gručenjem) za kar 14%. Povečini je bilo izboljšanje dokaj minimalno za nekaj odstotkov. Izboljšanje je bilo značilno v devetih primerih, poslabšanje pa samo v enem.

	brez zlivanja	večciljno regresijsko zlivanje	enociljno regresijsko zlivanje	enociljno zlivanje z gručenjem
TTP-LR	0.907	0.902	0.905	**0.91**
TTB-LR	**0.907**	0.902	0.906	0.906
TBO-LR	0.858	0.87	0.858	**0.889**
TBTP-LR	0.858	0.859	**0.895**	0.864
TPTB-LR	**0.869**	0.858	0.848	0.865
PTP-LR	**0.7**	0.673	0.692	0.668
SASD-LR	0.844	0.892	**0.908**	0.888
SASS-LR	0.844	0.908	**0.927**	0.789
TTP-NB	0.88	**0.884**	0.879	0.869
TTB-NB	0.88	**0.882**	0.881	0.875
TBO-NB	**0.865**	**0.865**	0.859	0.859
TBTP-NB	0.865	0.834	**0.871**	0.833
TPTB-NB	0.856	**0.863**	0.858	0.812
PTP-NB	0.649	0.651	0.661	**0.663**
SASD-NB	0.844	**0.952**	**0.952**	**0.952**
SASS-NB	0.844	0.955	**0.969**	0.886
TTP-DT	0.872	**0.876**	0.872	0.871
TTB-DT	0.872	0.862	0.872	**0.884**
TBO-DT	0.778	0.776	**0.782**	0.764
TBTP-DT	0.778	0.783	**0.826**	0.785
TPTB-DT	0.81	0.822	**0.839**	0.822
PTP-DT	0.635	0.612	0.607	**0.639**
SASD-DT	0.967	**0.983**	0.967	0.971
SASS-DT	0.967	**0.971**	0.927	0.952
TTP-KNN	0.868	0.873	**0.888**	0.868
TTB-KNN	0.868	0.878	**0.887**	0.876
TBO-KNN	0.827	0.808	**0.833**	0.809
TBTP-KNN	0.827	0.832	**0.871**	0.739
TPTB-KNN	0.841	0.839	**0.856**	0.837
PTP-KNN	**0.679**	0.667	0.677	0.668
SASD-KNN	0.877	**0.882**	0.877	0.811
SASS-KNN	**0.877**	0.842	0.844	0.743
TTP-RF	0.878	0.894	**0.898**	0.887

TTB-RF	0.878	0.882	**0.888**	0.881
TBO-RF	0.85	0.857	**0.869**	0.815
TBTP-RF	0.85	0.814	0.832	**0.85**
TPTB-RF	**0.848**	0.818	**0.848**	0.837
PTP-RF	0.644	0.658	**0.675**	0.647
SASD-RF	**0.983**	0.967	0.954	0.971
SASS-RF	**0.983**	0.971	0.967	0.911
TTP-SVM	0.899	**0.914**	0.908	0.897
TTB-SVM	0.899	**0.905**	**0.905**	0.896
TBO-SVM	0.845	**0.859**	0.839	0.858
TBTP-SVM	0.845	0.859	**0.877**	0.827
TPTB-SVM	0.856	**0.865**	0.856	0.839
PTP-SVM	0.666	0.673	**0.683**	0.67
SASD-SVM	**0.94**	0.904	0.89	0.852
SASS-SVM	**0.94**	0.892	**0.94**	0.805
TTP-NN	0.866	**0.899**	0.895	0.876
TTB-NN	0.866	**0.875**	0.873	0.874
TBO-NN	0.839	0.858	**0.87**	0.838
TBTP-NN	0.839	0.8	**0.875**	0.857
TPTB-NN	0.85	0.843	0.827	**0.852**
PTP-NN	**0.716**	0.675	0.666	0.703
SASD-NN	0.869	0.877	**0.894**	0.861
SASS-NN	0.869	0.861	**0.954**	0.824

Tabela 6.1: *Klasifikacijske točnosti vseh uporabljenih klasifikatorjev pri uporabi osnovne množice in zlitih množic. Z odebeljenim tiskom so napisane najvišje klasifikacijske točnosti za trojico: par množic - klasifikator - metoda zlivanja. Z rdečo barvo so označene klasifikacijske točnosti, ki so značilno drugačne od klasifikacijskih točnosti modela brez zlivanja.*

Povzetek rezultatov je prikazan v tabeli 6.2, ki prikazuje število značilnih izboljšav/poslabšanj klasifikacijske točnosti glede na metodo zlivanja in klasifikator. Vsako od metod zlivanja smo statistično primerjali z metodo brez zlivanja z Wilcoxonovim testom. Enociljno regresijsko zlivanje je v povprečju dosegalo najboljše rezultate, takoj za njim pa je bilo večciljno regresijsko zlivanje. Pri nobeni od teh metod ni bilo značilnega poslabšanja. Edino značilno poslabšanje je imelo enociljno zlivanje z gručenjem v kombinaciji s klasifikatorjem SVM. Klasifikator SVM je imel največje število značilnih razlik med metodami zlivanja in metodo brez zlivanja množic.

	večciljno regresijsko zlivanje	enociljno regresijsko zlivanje	enociljno zlivanje z gručenjem	skupaj
LR	-	-	-	-
NB	-	-	-	-
DT	-	-	-	-
KNN	-	2/0	-	2/0
RF	1/0	1/0	-	2/0
SVM	1/0	2/0	0/1	3/1
NN	1/0	1/0	-	2/0
skupaj	3/0	6/0	0/1	

Tabela 6.2: *Število značilnih izboljšav/poslabšanj klasifikacijske točnosti glede na metodo zlivanja in klasifikator.*

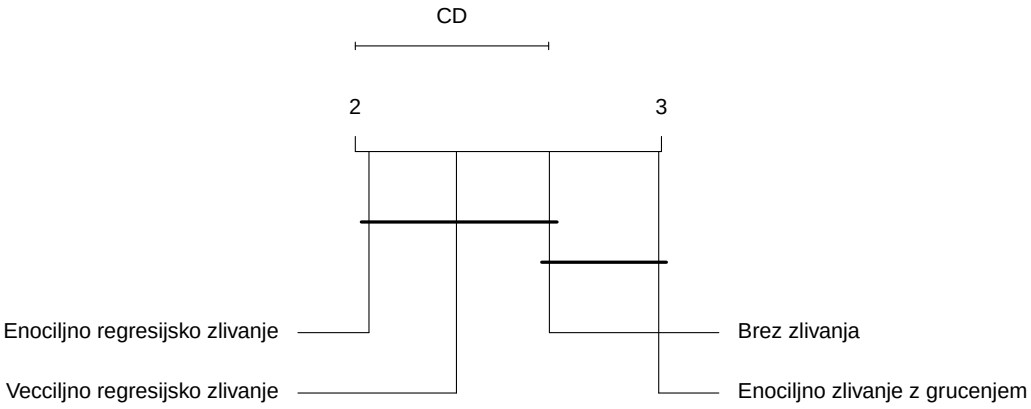
Vse štiri metode smo tudi statistično primerjali s Friedmanovim testom, kjer smo dobili p-vrednost 0.0008383. To pomeni, da se klasifikacijske točnosti metod med seboj značilno razlikujejo in da lahko ničelno hipotezo zavržemo.

Podrobnejša statistična primerjava posameznih metod s post-hoc testom Nemenyi-a je pokazala, da nobena od metod zlivanja ni značilno boljša od metode brez zlivanja podatkov. So pa bile razlike med predlaganimi metodami zlivanja. Kot lahko vidimo na grafu rangov na sliki 6.1, sta bili enociljno in večciljno regresijsko zlivanje značilno boljša od enociljnega zlivanja z gručenjem. Prav tako je razvidno, da se je to zlivanje v povprečju najslabše odrezalo in je bilo celo slabše od modelov brez uporabe zlivanja.

Statistično primerjavo metod za zlivanje smo naredili tudi ločeno za vsak posamezni klasifikator. Rezultati statističnih primerjav so prikazani na sliki 6.2. Značilno različne so bile metode samo pri dveh klasifikatorjih. Tudi tu nobena metoda zlivanja ni bila boljša od metode brez zlivanja podatkov. Enociljno zlivanje z gručenjem se je pri obeh odrezalo najslabše. Bilo je značilno slabše od večciljnega regresijskega zlivanja pri klasifikatorju SVM in enociljnega regresijskega zlivanja pri klasifikatorju KNN.

6.2 Uspešnost posameznih klasifikatorjev

Statistično primerjavo smo naredili tudi za posamezne klasifikatorje preko vseh uporabljenih množic (osnovnih in zlitih). Rezultat Friedmanovega testa je bila p-vrednost 4.6×10^{-8}. Klasifikacijske točnosti klasifikatorjev se med seboj značilno razlikujejo in ničelno hipotezo



Slika 6.1: *Prikaz uspešnosti vseh uporabljenih metod.*

lahko zavržemo.

S post-hoc testom Nemenyi-a smo podrobneje statistično primerjali posamezne klasifikatorje in ugotovili:

1. linearna regresija je značilno boljša od nevronskih mrež, metode k-najbližjih sosedov in odločitvenega drevesa, ter

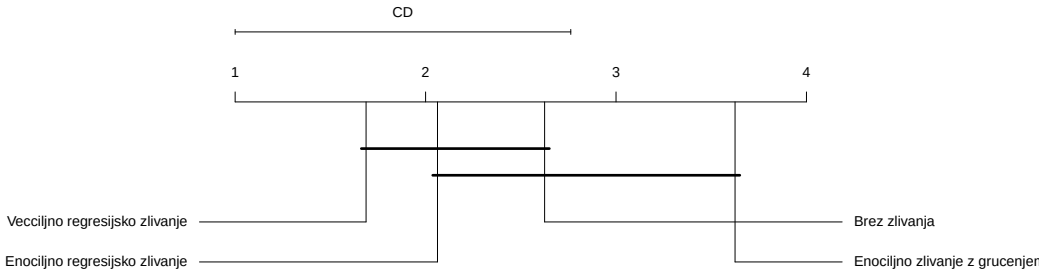
2. naključni gozdovi in metoda podpornih vektorjev sta značilno boljši od metode k-najbližjih sosedov in odločitvenega drevesa.

Primerjavo klasifikatorjev lahko vidimo na grafu rangov na sliki 6.3. V povprečju se je najbolje odrezala linearna regresija, najslabše pa metoda k-najbližjih sosedov.

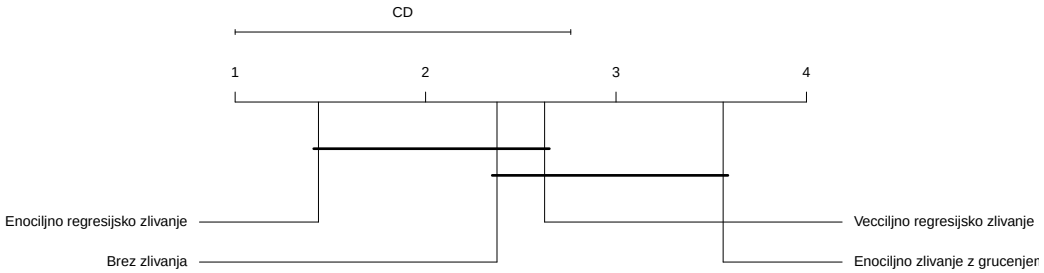
Prav tako smo naredili statistično primerjavo tudi ločeno za vsako metodo zlivanja – brez zlivanja (referenčna metoda), zlivanje z večciljnimi modeli, zlivanje z enociljnimi regresijskimi modeli in zlivanje z gručenjem. Rezultati statističnih primerjav so razvidni iz grafov na sliki 6.4. Najbolj konsistentna linearna regresija, ki je pri treh od štirih metod je dosegla najvišji rang. Samo pri uporabi množic, zlitih z enociljnim zlivanjem z gručenjem, ni dosegla najvišjega ranga. Pravo nasprotje je odločitveno drevo, ki samo pri uporabi množic zlitih z enociljnim zlivanjem z gručenjem, ni padel na zadnje mesto.

Slika 6.4a prikazuje uspešnost klasifikatorjev na osnovnih množicah. Najboljša povprečna ranga sta dosegla linearna regresija in metoda podpornih vektorjev, ki sta bila tudi značilno boljša od vseh preostalih klasifikatorjev.

Uspešnost klasifikatorjev pri uporabi množic zlitih z večciljnim regresijskim zlivanjem prikazuje slika 6.4b. Linearna regresija je bila značilno boljša od metode k-najbližjih



(a) Klasifikator SVM.



(b) Klasifikator KNN.

Slika 6.2: *Prikaz uspešnosti vseh uporabljenih metod glede na klasifikator.*

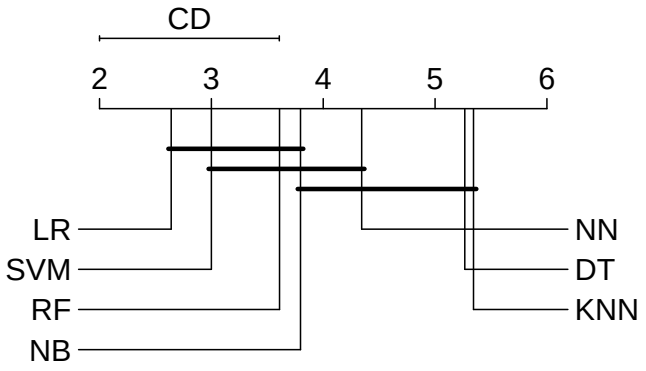
sosedov, naivni Bayes in metoda podpornih vektorjev pa sta bila značilno boljša od odločitvenega drevesa.

Na sliki 6.4c lahko vidimo uspešnost klasifikatorjev na množicah, zlitih z enociljnim regresijskim zlivanjem. Linearna regresija, metoda podpornih vektorjev in nevronske mreže so bile glede na klasifikacijsko točnost značilno različne samo od odločitvenega drevesa.

Zadnja slika 6.4d prikazuje spešnost klasifikatorjev pri uporabi množic zlitih z enociljnim zlivanjem z gručenjem. Klasifikatorja, ki temeljita na grajenju dreves, sta bila tu v povprečju najboljša. Naključna drevesa so bila tu značilno boljša od nevronske mreže, linearne regresije in metode k-najbližjih sosedov, odločitveno drevo pa je bilo značilno boljše samo od k-najbližjih sosedov.

6.3 Diskusija

Razlogov, zakaj zlivanje v povprečju ni dosegalo značilno boljših rezultatov, je možnih več. Iz tabele 5.1 je razvidno, da imajo množice s podatki o proteinih in podatki, izlučenimi iz slik MRI, izjemno veliko število atributov. Preostale množice pa vsebujejo predvsem manjše število diskretnih atributov z velikimi zalogami vrednosti. To pomeni, da se v času zlivanja število atributov teh množic napihne. Naša odločitev je bila, da vedno dodamo



Slika 6.3: *Prikaz uspešnosti vseh uporabljenih klasifikatorjev.*

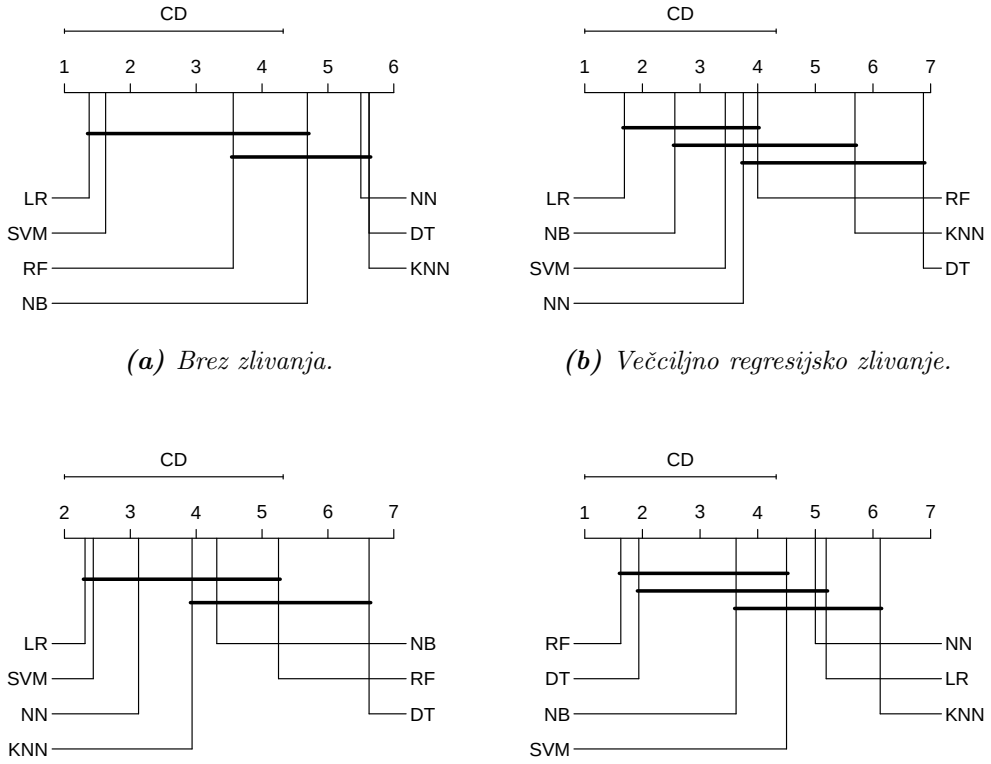
50 atributov v procesu zlivanja. V primerjavi z obstoječimi atributi osnovne množice je to le manjši del. Velik del dodanih atributov je zato pred učenjem klasifikatorja odpadel (ko smo naredili izbor atributov) in imel posledično manjši vpliv na končni model.

Na enociljno zlivanje z gručenjem je verjetno vplivala izbira parametrov. Število skupin pri metodi k voditeljev je bilo vedno enako, ne glede na učno množico. To pomeni, da je bil regresor, ki je uporabljal to metodo za osnovo, različno natančen pri napovedovanju vrednosti manjkajočih atributov. Gručenje je na nek način diskretiziralo zalogo vrednosti teh atributov. Kvaliteta diskretizacije pa je bila odvisna od kvalitete samih atributov in števila gruč.

Število atributov in primerov v posamezni množici je prav tako vplivalo na celoten proces napovedovanja. V procesu zlivanja smo skupne atribute selektivno izbrali z algoritmom RReliefF, ki ima probleme z ocenjevanjem atributov, ko je število pomembnih atributov relativno majhno v primerjavi s številom nepomembnih atributov. Še večji problem se pojavi, če je še število primerov majhno. Za množice z atributi, ki opisujejo proteine in lastnosti možganov, je to zelo velik problem. Te množice so imele med 200 in 500 primerov in med 1000 in 3000 atributov. Poleg tega smo algoritem ReliefF uporabili za izbiro atributov pred grajenjem modelov, kar pomeni, da obstaja možnost, da so bili atributi dvakrat napačno ocenjeni in da je to vplivalo na zgrajeni model.

Na napovedno točnost so vplivale tudi modalnosti podatkov v množici, uporabljeni za grajenje modela. Višje napovedne točnosti so dosegale množice s podatki o rezultatih različnih testov. Pri množicah, ki teh modalnosti niso vsebovale (npr. Proteins), so bile napovedne točnosti med 10-20% nižje kot pri ostalih.

Poleg tega velja tudi omeniti problem pripisane diagnoze posameznih primerov v uporabljenih množicah. Kljub temu, da so diagnozo dodelili strokovnjaki, to ni nujno pravilna diagnoza. Pravilno diagnozo je mogoče dobiti šele ob smrti pacienta. Nepravilne oznake



(a) *Brez zlivanja.*

(b) *Večciljno regresijsko zlivanje.*

(c) *Enociljno regresijsko zlivanje.*

(d) *Enociljno zlivanje z gručenjem.*

Slika 6.4: *Prikaz uspešnosti vseh uporabljenih klasifikatorjev glede na metodo zlivanja.*

primerov negativno vplivajo tako na učenje napovednega modela z učnim delom množice kot na preverjanje kvalitete tega modela z uporabo testnega dela množice. Vse to pa ima vpliv na napovedno točnost. Ta problem se pojavlja tudi v literaturi. Najbolj viden je bil ta problem v metodi, ki so jo predlagali v [25].

Presenetljiva je uspešnost linearne regresije. Zanjo je znano, da ima probleme z velikim številom kategoričnih atributov in nelinearnimi atributi. Taki atributi so bili v naših množicah prisotni – atributi iz rezultatov kognitivnih testov so bili povečini kategorični, atributi iz slik MRI pa nelinearni. V literaturi se logistična regresija sicer pojavlja, vendar precej redkeje kot SVM ali naključni gozdovi. Mi smo zasledili njeno uporabo v samo nekaj primerih, pri čemer so v večini teh primerov raziskovalci uporabljali samo rezultate kognitivnih testov. V naši nalogi je bila enakovredna naključnim gozdovom, SVM, naivnemu Bayesu in nevronski mreži.

Poglavje 7

Zaključek

V okviru magistrske naloge smo obravnavali problem napovedovanja Alzheimerjeve bolezni. Pregled literature je pokazal, da obstaja veliko število študij, kjer so zbrani raznoliki podatki. Večina teh študij je zasebnih in njihovi podatki niso javno dostopni. Preostale, ki pa so javne, ciljajo predvsem na raziskovalce z vseh področij, ki jih zanima čimprejšnje odkritje vzrokov za pojavitev Alzheimerjeve bolezni in zdravil za blaženje posledic oziroma celo ozdravitev. Kljub temu, da so na voljo različni tipi podatkov, se raziskovalci povečinoma odločajo za uporabo le enega. V preteklosti so raziskovalci z drugih medicinskih področij že dokazali, da pametna uporaba več tipov podatkov lahko veliko doprinese k razumevanju problema in napovedovanju bolezni.

V naši nalogi smo se lotili napovedovanja Alzheimerjeve bolezni z uporabo zlivanja različnih tipov podatkov, dobljenih iz različnih virov. Implementirali smo metodo zgodnje integracije atributov. Preko skupnih atributov dveh množic smo prvi množici dodali manjkajoče atribute iz druge. Modele, naučene na zlitih množicah, smo primerjali z modeli, zgrajenimi na originalnih množicah.

Za čimbolj realistično primerjavo in evalvacijo modelov za zlivanje smo zbrali 8 različnih množic podatkov iz treh različnih virov. Naš cilj je bil pridobiti čimbolj različne množice za kvalitetno ocenitev predlagane metode zlivanja. Podatki so bili zato različnih modalnosti, množice pa so imele različno število tako atributov kot primerov. Podatki iz zbirke OASIS so bili celo v obliki slik MRI, ki smo jih obdelali in iz njih izvlekli različne deskriptivne atribute.

Rezultati so pokazali, da noben od modelov, zgrajenih na podatkih, zlitih s predlaganimi metodami, ni bil statistično značilno boljši od modelov, zgrajenih na osnovnih množicah. Kljub temu pa so predlagane metode v nekaterih primerih izboljšale napovedno točnost tudi do 14%. Dve izmed treh metod sta imeli tudi višji povprečni rang od osnovne metode. Za nekatere množice se je izkazalo, da je bolje zamenjati napovedni mo-

del kot pa strategijo zlivanja, kar pomeni, da dodani atributi bodisi niso doprinesli nobene kvalitetne informacije ali pa da jih je bilo v primerjavi z originalnimi premalo in so se porazgubili med ocenjevanjem atributov za grajenje modela.

Uspeh pri nekaterih poskusih vendarle nakazuje na skriti potencial, ki ga ima zlivanje. Izboljšave so možne v celotnem cevovodu grajenja modela. Za dodajanje atributov smo v nalogi uporabili klasične metode za nadzorovano učenje in gručenje. Zanimivo bi bilo preiskusiti, kakšen drug način, na primer, s povezovalnimi pravili (Apriori algoritem).

Poleg tega bi lahko uporabili druge algoritme za izbiro atributov. Algoritmi iz družine Relief imajo velike probleme z določitvijo pomembnih atributov, če je število nepomembnih v množici veliko. Poleg tega potrebujejo veliko število primerov za dobro oceno kakovosti. Ta dva problema sta se pri nas pojavila v kar nekaj množicah.

Zanimivo bi bilo tudi raziskati vpliv različnih parametrov algoritmov strojnega učenja, predvsem pri zlivanju podatkov. V to spadajo tako preprosti parametri za zmanjševanje prilagajanja modela podatkom kot tudi bolj kompleksni, kot je na primer struktura nevronske mreže. V tej nalogi se tega problema nismo lotili, vendar lahko na ta način verjetno dodatno izboljšamo kvaliteto dodanih atributov.

Literatura

[1] International Statistical Classification of Diseases and Related Health Problems 10th Revision, http://apps.who.int/classifications/icd10/browse/2016/en#/F00, obiskano: 10.9.2017.

[2] Dementia, http://www.who.int/mediacentre/factsheets/fs362/en/, obiskano: 10.9.2017.

[3] M. Gendron, Skrivnost, imenovana Alzheimer, Chiara, 2015.

[4] I. Grant, K. M. Adams, Neuropsychological assessment of neuropsychiatric and neuromedical disorders, Oxford University Press, 2009.

[5] D. Andoljšek, M. Reingold, R. Berkow, M. H. Beers, A. J. Fletcher, et al., Veliki zdravstveni priročnik, Mladinska knjiga, 2005.

[6] P. A. DeFina, R. S. Moser, M. Glenn, J. D. Lichtenstein, J. Fellus, Alzheimer's disease clinical and research update for health care practitioners, Journal of aging research 2013.

[7] FreeSurfer, https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki, obiskano: 10.9.2017.

[8] A. Klein, S. S. Ghosh, F. S. Bao, J. Giard, Y. Hame, E. Stavsky, N. Lee, B. Rossa, M. Reuter, E. Chaibub Neto, A. Keshavan, Mindboggling morphometry of human brains, bioRxivarXiv:https://www.biorxiv.org/content/early/2016/12/03/091322.full.pdf, doi:10.1101/091322.
URL https://www.biorxiv.org/content/early/2016/12/03/091322

[9] N. J. Tustison, P. A. Cook, A. Klein, G. Song, S. R. Das, J. T. Duda, B. M. Kandel, N. van Strien, J. R. Stone, J. C. Gee, et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, Neuroimage 99 (2014) 166–179.

[10] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin, A. D. N. Initiative, et al., Does feature selection improve classification accuracy? impact of sample size and feature

selection on classification using anatomical magnetic resonance images, Neuroimage 60 (1) (2012) 59–70.

[11] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1) (2002) 389–422. doi:10.1023/A:1012487302797.
URL https://doi.org/10.1023/A:1012487302797

[12] M. Liu, D. Zhang, D. Shen, A. D. N. Initiative, et al., Ensemble sparse classification of Alzheimer's disease, NeuroImage 60 (2) (2012) 1106–1116.

[13] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, S. C. Johnson, A. D. N. Initiative, et al., Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset, Neuroimage 48 (1) (2009) 138–149.

[14] E. Hosseini-Asl, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by adaptation of 3D convolutional network, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 126–130.

[15] M. López, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, R. Chaves, P. Padilla, M. Gómez-Río, A. D. N. Initiative, et al., Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease, Neurocomputing 74 (8) (2011) 1260–1271.

[16] J. S. Kippenhan, W. W. Barker, S. Pascal, J. H. Nagel, R. Duara, Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects.

[17] Y. Cho, J.-K. Seong, Y. Jeong, S. Y. Shin, A. D. N. Initiative, et al., Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data, Neuroimage 59 (3) (2012) 2217–2230.

[18] K. R. Gray, R. Wolz, R. A. Heckemann, P. Aljabar, A. Hammers, D. Rueckert, A. D. N. Initiative, et al., Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease, Neuroimage 60 (1) (2012) 221–229.

[19] M. Huang, W. Yang, Q. Feng, W. Chen, A. D. N. Initiative, et al., Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease, Scientific reports 7.

[20] J. D. Doecke, S. M. Laws, N. G. Faux, W. Wilson, S. C. Burnham, C.-P. Lam, A. Mondal, J. Bedo, A. I. Bush, B. Brown, et al., Blood-based protein biomarkers for diagnosis of Alzheimer disease, Archives of neurology 69 (10) (2012) 1318–1325.

[21] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, D. R. Galasko, M. Jutel, A. Karydas, et al., Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins., Nature medicine 13 (11).

[22] D. A. Llano, V. Devanarayan, A. J. Simon, A. D. N. I. (ADNI, et al., Evaluation of plasma proteomic data for Alzheimer disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer disease, Alzheimer Disease & Associated Disorders 27 (3) (2013) 233–243.

[23] W. S. Pritchard, D. W. Duke, K. L. Coburn, N. C. Moore, K. A. Tucker, M. W. Jann, R. M. Hostetler, EEG-based, neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by non-linear EEG measures, Electroencephalography and clinical Neurophysiology 91 (2) (1994) 118–130.

[24] E. Challis, P. Hurley, L. Serra, M. Bozzali, S. Oliver, M. Cercignani, Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI, NeuroImage 112 (2015) 232–243.

[25] T. Galili, A. Mitelpunkt, N. Shachar, M. Marcus-Kalish, Y. Benjamini, Categorize, cluster, and classify: a 3-c strategy for scientific discovery in the medical informatics platform of the human brain project, in: International Conference on Discovery Science, Springer, 2014, pp. 73–86.

[26] M. Tierney, J. Szalai, W. Snow, R. Fisher, A. Nores, G. Nadon, E. Dunn, P. S. George-Hyslop, Prediction of probable Alzheimer's disease in memory-impaired patients A prospective longitudinal study, Neurology 46 (3) (1996) 661–665.

[27] J. J. Gomar, M. T. Bobes-Bascaran, C. Conejero-Goldberg, P. Davies, T. E. Goldberg, A. D. N. Initiative, et al., Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative, Archives of general psychiatry 68 (9) (2011) 961–969.

[28] M. Ewers, C. Walsh, J. Q. Trojanowski, L. M. Shaw, R. C. Petersen, C. R. Jack, H. H. Feldman, A. L. Bokde, G. E. Alexander, P. Scheltens, et al., Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance, Neurobiology of aging 33 (7) (2012) 1203–1214.

[29] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, et al., Multimodal classification of Alzheimer's disease and mild cognitive impairment, Neuroimage 55 (3) (2011) 856–867.

[30] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, A. D. N. Initiative, et al., Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, NeuroImage 65 (2013) 167–175.

[31] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage 101 (2014) 569–582.

[32] R. Polikar, A. Topalis, D. Parikh, D. Green, J. Frymiare, J. Kounios, C. M. Clark, An ensemble based data fusion approach for early diagnosis of Alzheimer's disease, Information Fusion 9 (1) (2008) 83–95.

[33] H. Zhang, The optimality of naive Bayes, AA 1 (2) (2004) 3.

[34] I. Kononenko, M. Šikonja, Inteligentni sistemi, Založba FE in FRI, 2010.
URL `https://books.google.si/books?id=LxZDMwEACAAJ`

[35] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and computing 14 (3) (2004) 199–222.

[36] G. Čepin, Večciljno učenje v klasifikaciji in regresiji.
URL `https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=81190`

[37] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine learning 53 (1-2) (2003) 23–69.

[38] M. Robnik-Šikonja, I. Kononenko, An adaptation of Relief for attribute estimation in regression, in: Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 1997, pp. 296–304.

[39] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Advances in neural information processing systems, 2006, pp. 507–514.

[40] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE transactions on pattern analysis and machine intelligence 37 (1) (2015) 41–53.

[41] P. Pavlidis, J. Weston, J. Cai, W. S. Noble, Learning gene functional classifications from multiple data types, Journal of computational biology 9 (2) (2002) 401–411.

[42] The AddNeuroMed Study, `https://www.synapse.org/#!Synapse:syn2790911/wiki/235387`, obiskano: 10.9.2017.

[43] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L.-O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Mecocci, M. Tsolaki, et al., Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups—the AddNeuroMed study, Neurobiology of Aging 32 (7) (2011) 1198–1206.

[44] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, et al., The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's & dementia 7 (3) (2011) 263–269.

[45] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, Journal of cognitive neuroscience 19 (9) (2007) 1498–1507.

[46] MRI, `https://www.synapse.org/#!Synapse:syn2795021`, obiskano: 10.9.2017.

[47] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine learning research 7 (Jan) (2006) 1–30.