Učenje Bayesovih mreža iz cenzuriranih podataka o preživljavanju

Štajduhar, Ivan

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Ivan Štajduhar

UČENJE BAYESOVIH MREŽA IZ CENZURIRANIH PODATAKA O PREŽIVLJENJU

DOKTORSKA DISERTACIJA

Zagreb, 2010.

Doktorska disertacija izrađena je na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu i Zavodu za računarstvo Tehničkog fakulteta Sveučilišta u Rijeci.

Mentor: Bojana Dalbelo Bašić

Broj stranica: 173

Disertacija broj:

Povjerenstvo za ocjenu doktorske disertacije:

1. Dr. sc. Nikola Bogunović, redoviti profesor
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
2. Dr. sc. Bojana Dalbelo Bašić, redovita profesorica
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
3. Dr. sc. Dragan Gamberger, znanstveni savjetnik
   Institut Ruđer Bošković, Zagreb

Povjerenstvo za obranu doktorske disertacije:

1. Dr. sc. Nikola Bogunović, redoviti profesor
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
2. Dr. sc. Bojana Dalbelo Bašić, redovita profesorica
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
3. Dr. sc. Dragan Gamberger, znanstveni savjetnik
   Institut Ruđer Bošković, Zagreb
4. Dr. sc. Tomislav Šmuc, viši znanstveni suradnik
   Institut Ruđer Bošković, Zagreb
5. Dr. sc. Siniša Šegvić, docent
   Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

Datum obrane disertacije: 26. ožujka 2010. godine

Sadržaj

Poglavlje 1

Uvod

Analiza preživljenja je zajednički naziv za skup statističkih postupaka za analiziranje podataka, kod kojih je varijata (engl. *variate*, varijabla) od interesa vrijeme koje je potrebno da se neki događaj zbije. Modeliranje preživljenja znači utvrđivanje pravila procjene distribucije vremena do događaja (ishoda) za neki objekt, temeljeno na njegovim karakteristikama (Kleinbaum, 2005). U medicini je uobičajeno modeliranje preživljenja pacijenata te je događaj koji se prati obično razvoj neke bolesti, reakcija na terapiju, povrat bolesti ili smrt (Lee & Wang, 2003). Korištenjem popularnih statističkih postupaka, poput regresijskog modela proporcionalnih hazarda (Cox, 1972), klasifikacijskih i regresijskih stabala (Breiman *et al.*, 1984), regresije hazarda (Kooperberg *et al.*, 1995) ili Bayesovih stabala (Clarke & West, 2008), moguće je odrediti modele preživljenja iz podataka. Takvi modeli moći će predvidjeti krivulje preživljenja, temeljene na dokazima, to jest vrijednostima opaženih kovarijata.

Iako su sposobni modelirati nelinearne odnose i stvarati tumačljive modele, postupci strojnog učenja (Duda *et al.*, 2001) rijetko se koriste u analizi preživljenja. Razlog tome je da oni, *per se*, nisu sposobni predvidjeti krivulju preživljenja, to jest vrijeme do zbivanja ishoda. Umjesto toga, postupci strojnog učenja sposobni su predvidjeti hoće li se ili ne ishod zbiti, sukladno dokazima, neovisno o vremenu. Takva predviđanja mogu u nekim slučajevima biti iznimno korisna. Npr. ako bi liječnik onkolog trebao odlučiti hoće li propisati postoperativnu kemoterapiju za pacijenticu oboljelu od raka dojke, mogao bi usporediti obje predviđene vjerojatnosti povrata raka temeljene na njenim karakteristikama, prema dokaznim scenarijima (s ili bez kemoterapije), i sukladno tome donijeti svoju odluku. U svim slučajevima u kojima vrijeme do zbivanja ishoda nije ključno, svi se postupci strojnog učenja mogu uporabiti za učenje medicinskih prognostičkih modela iz podataka (Lucas & Abu-Hanna, 1999). Problem se pojavljuje kad se ti postupci pokušaju iskoristiti za učenje iz cenzuriranih podataka. Tema ove disertacije jest prilagodba nekih postupaka strojnog učenja radu s cenzuriranim podacima o preživljenju.

Bayesove mreže (Pearl, 1988) su grafički prikaz distribucija vjerojatnosti. Uglavnom se koriste za prirodno i lako razumljivo predstavljanje neodređenosti u raznim domenama. Bayesova mreža sastoji se od usmjerenog acikličkog grafa (engl. *direc-*

ted acyclic graph, DAG) i skupa tablica uvjetnih vjerojatnosti. Oni zajedno opisuju inherentnu distribuciju vjerojatnosti u domeni (Pearl, 2000). Struktura mreže je kvalitativni dio modela jer usmjerenim lukovima prikazuje interakcije kovarijata, u smislu uzročno-posljedičnih veza, te odgovara na pitanje o međusobnoj ovisnosti kovarijata. Skup tablica uvjetnih vjerojatnosti je kvantitativni dio Bayesove mreže jer opisuje kako kovarijate, koje su međusobno povezane, ovise jedne o drugima (kroz uvjetne vjerojatnosti). Grafički prikaz Bayesove mreže omogućuje nam jednostavnu interpretaciju uzročnosti slijedom strelica na lukovima. Zaključivanje iz vjerojatnosti slijedi strukturu mreže i koristi se pri klasifikaciji. Iz razloga što je svaka kovarijata u Bayesovoj mreži neovisna od svojih ne-potomaka, ako su poznati njeni roditelji, zaključivanje je računski brzo.

Bayesove mreže mogu se koristiti za predstavljanje uzročnih utjecaja u probabilistički izraženim interakcijama kovarijata. Takav model predstavljanja odgovara ljudskom zaključivanju o uzročnosti i neizvjesnosti. Pored toga, mogu se naučiti iz podataka. Bayesove mreže su, iz navedenih razloga, izvrstan alat za predstavljanje znanja. U novije vrijeme uživaju sve veću popularnost u biomedicini i zdravstvenoj skrbi (Lucas *et al.*, 2004) za potrebe dijagnoze, tretmana, prognoze i otkrivanja funkcijskih interakcija. Uspostavljeni su različiti modeli u području onkologije (Lucas *et al.*, 1998), zaraznih bolesti (Andreassen *et al.*, 1999; Lucas *et al.*, 2000) i transplantacije (Hoot & Aronsky, 2005).

U posljednjih 15 godina nastalo je pregršt postupaka za učenje Bayesovih mreža iz podataka. Ti postupci učinkovito rukuju kako učenjem strukture, tako i ostalih parametara mreže iz potpunih (Cooper & Herskovits, 1992; Heckerman *et al.*, 1995; Lam & Bacchus, 1994) i nepotpunih podataka (Friedman, 1998), to jest onih koji imaju nedostajuće vrijednosti. Iz naučene strukture mreže moguće je steći novo znanje o mogućim uzročnim odnosima prisutnim u domeni (Pearl, 1988). Na utvrđenoj (naučenoj) Bayesovoj mreži je jednostavno mjerljive veličine moguće iskoristiti za određivanje onih teže mjerljivih.

1.1 Svrha i pregled doktorske disertacije

U ovoj su disertaciji predstavljena moguća rješenja problema učenja prognostičkih Bayesovih mreža iz cenzuriranih podataka o preživljenju, prvenstveno za potrebe klasifikacije (predviđanja konačnog ishoda), no isto tako i poimanja i razumijevanja interakcija kovarijata. Bayesove su mreže, kao i mnoge druge metode strojnog učenja, također sposobne modelirati nelinearnosti u domenama, ali su još zanimljivije upravo zbog jednostavnosti predstavljanja interakcija u domeni. Jednostavnost interpretacije i razumljivost ključna su motivacija za primjenu takvih modela u potpori odlučivanju, i komparativna prednost pred statističkim modelima spomenutim ranije.

U nastavku je opisana struktura disertacije. U poglavlju 2 predstavljeni su temeljni koncepti korišteni u radu (Bayesove mreže i analiza preživljenja). U sekciji 2.4 dan je kratak pregled srodnih radova. U poglavlju 3 predstavljeni su neki od popularnih postupaka prilagodbe podataka o preživljenju za algoritme strojnog učenja. Poglavlje

4 sadrži detaljan opis korištenih algoritama za učenje Bayesovih mreža te opis naivnog Bayesovog klasifikatora i Coxove regresije. Način i metrike kojima su postupci testirani predstavljeni su u poglavlju 5. Postupci su prvo detaljno testirani u simulacijskoj studiji (poglavlje 6), uspoređujući prvo kvalitetu performansi naučenih modela u zaključivanju, odnosno klasifikaciji (sekcija 6.1), a zatim i sposobnost učenja ispravnih struktura mreža (sekcija 6.2). Postupci su dodatno testirani na tri realne domene iz kliničke medicine u poglavlju 7. U poglavlju 8 nalazi se interpretacija dobivenih rezultata te zaključak.

Doprinos ove disertacije je postupak prilagodbe podataka o preživljenju njihovom pripremnom obradom odstranjivanjem šuma cenzure, temeljen na izglednosti podataka u Bayesovim mrežama (sekcija 3.5). Po primjeni tog postupka, nad podacima je moguće upotrijebiti bilo koji postupak standardnog strojnog učenja (pa tako i standardnih algoritama za učenje Bayesovih mreža) za učenje smislenog modela. Rezultati predstavljeni u simulacijskoj studiji (sekcija 6.1.4) pokazuju očitu superiornost tog postupka u klasifikaciji nad svim ostalim poznatim postupcima. Doprinos je još i temeljita usporedba različitih pristupa rukovanju cenzuriranim podacima o preživljenju u strojnom učenju detaljno popraćena prikladnim statističkim testovima, kako na umjetno generiranim podacima (sekcije 6.1.4 i 6.2.3), tako i na tri dobro poznate i javno dostupne realne domene iz svijeta kliničke medicine (poglavlje 7). Na jednom su mjestu prikazane i topologije Bayesovih mreža (dodatak D) naučene primjenom različitih pristupa rukovanju cenzuriranim podacima o preživljenju za te realne domene. Jedan od korištenih pristupa rukovanju cenzuriranim podacima o preživljenju je i izvedba postupka podvajanja cenzuriranih primjeraka uz težinske faktore (sekcija 3.4) za učenje Bayesovih mreža.

1.2 Korišteni alati

U doktorskoj disertaciji je za izradu primjera i testiranje postupaka korištena sljedeća programska oprema: alat otvorenog koda za strojno učenje Weka (Witten & Frank, 2005), alat otvorenog koda za statističko računanje R (R Development Core Team, 2008) i programski jezik i okruženje MATLAB (verzija 7.5.0.342, R2007b). U izradi pisanog rada korišten je LaTeX 2_ε, dok su ilustracije stvorene alatom Microsoft Office Visio 2003 i alatom otvorenog koda Graphviz (Ellson *et al.*, 2002).

Poglavlje 2

Bayesove mreže i analiza preživljenja

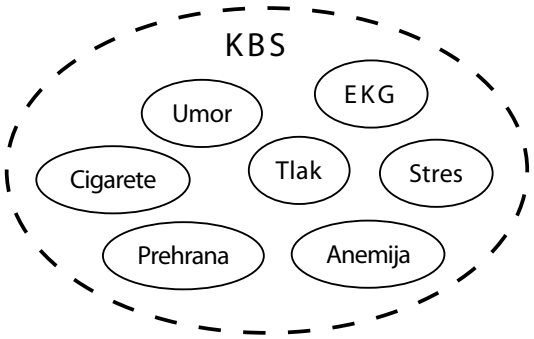
U ovom su poglavlju predstavljene Bayesove mreže, kao model predstavljanja stečenoga znanja, te analiza preživljenja, kao grana statistike koja se bavi analizom pojavljivanja nekog događaja od interesa (ispada) u biološkim i mehaničkim sustavima.

2.1 Primjer iz domene koronarne bolesti srca

Radi zornijeg prikaza opisanih koncepata, osmišljen je jednostavan pokazni model. Modelom je opisano međudjelovanje mogućih uzroka i simptoma koronarne bolesti srca, što je jedan dobro poznat klinički problem, lako razumljiv i medicinskim laicima. Potrebno je naglasiti da osmišljeni model nikako nije utemeljen na rezultatima znanstvenih istraživanja u kliničkoj medicini. Iako je model intuitivan, svoj je konačan oblik poprimio isključivo iz razloga kako bi se njime dali ilustrirati temeljni koncepti opisani u ovom radu.

Koronarna bolest srca (Braunwald *et al.*, 2001) najčešći je oblik bolesti koje pogađaju srce. Bolest je prisutna kada koronarni protok ne uspijeva zadovoljiti potrebe srčanog mišića i obližnjeg tkiva. Najčešći je uzročnik preuranjene smrti kod ljudi u ekonomski razvijenijim zemljama svijeta (Sjedinjene Američke Države, zemlje Europe, itd.), pretežno zbog prevladavajućih životnih navika.

U pokaznom su modelu (slika 2.1) odabrani neki od mogućih uzroka i simptoma koronarne bolesti srca (*KBS*) te neki od njima srodnih uzroka ili simptoma. To su: prekomjerna tjelesna masa na koju neposredno utječu loše prehrambene navike (*prehrana*) i nedostatak tjelovježbe, zatim uživanje nikotina (*cigarete*), svakodnevni stres (*stres*), povišeni krvni tlak (*tlak*), abnormalna slika elektrokardiograma (*EKG*), kronični umor (*umor*) te manjak željeza u krvi (*anemija*). Navedeni uzroci ili simptomi, kao i sama bolest, zajedno predstavljaju domenu, to jest skup oblika podataka kojima se može modelirati područje koronarne bolesti srca. U stvarnosti je broj uzroka i simptoma ove bolesti znatno veći; oni ovdje namjerno nisu bili korišteni, jer bi dodatno otežali razumijevanje za sada relativno jednostavnog modela. Detaljniji opis modela nalazi se



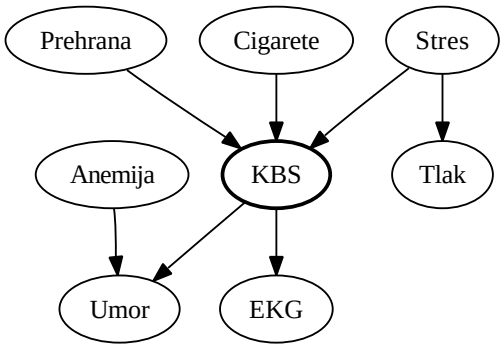
Slika 2.1: Početni model (domena) koronarne bolesti srca: mogući uzroci i posljedice.

u dodatku A.

Prekomjerna tjelesna masa (*prehrana*), uživanje nikotina (*cigarete*) i *stres* očiti su uzroci pojave koronarne bolesti srca. Povišeni krvni tlak (*tlak*), abnormalna slika elektrokardigorama (*EKG*) i prisutnost kroničnog umora (*umor*) očiti su simptomi te bolesti. No, povišeni krvni tlak (*tlak*) može se pojaviti i kao neposredna posljedica *stresa*, uz sasvim zdravo srce. Slično, kronični umor (*umor*) može biti posljedica neke sasvim druge bolesti, npr. manjka željeza u krvi (*anemija*).

Svaki od navedenih uzroka i posljedica bolesti, kao i podatak o prisutnosti same bolesti, predstavlja se slučajnom varijatom X_i, čiji je indeks predstavljen početnim slovom pojma koji opisuje, tj. $i \in \{P, C, S, T, U, E, A, K\}$. Skup varijata koje opisuju ovaj problem je dakle, $\mathbf{X} = \{X_P, X_C, X_S, X_T, X_U, X_E, X_A, X_K\}$. Varijate (varijable), za koje se pretpostavlja da ih je moguće koristiti za predviđanje nepoznate vrijednosti neke ciljne varijate, zovu se kovarijate (engl. *covariate, explanatory variable, independent variable, predictor variable*). Ciljna varijata, to jest varijata koja se predviđa, često se naziva i varijata od interesa (engl. *variate of interest*), ishod (engl. *outcome*) ili razred (engl. *class*), ovisno o znanstvenom području u kojem se primjenjuje, ali i o mjernoj ljestvici (regresija ili klasifikacija). Istraživanje provedeno ovom disertacijom pokriva isključivo klasifikaciju. U teoriji svaka kovarijata može biti i varijata od interesa, no u praksi su kovarijate obično one lako dostupne i jednostavno mjerljive veličine, dok je varijata od interesa ona nedostupna ili teško mjerljiva veličina koja se želi predvidjeti. U dijelovima ove disertacije u kojima je potrebno jasno istaknuti varijatu od interesa u skupu mogućih kovarijata, ona se označava slovom O (kao *outcome*). Sukladno tome, na primjeru modela koronarne bolesti srca vrijedi $\mathbf{X} = \{X_P, X_C, X_S, X_T, X_U, X_E, X_A\}$ i $O = X_K$. Moguću sposobnost predviđanja ishoda u odnosu na poznate vrijednosti kovarijata pišemo $\mathbf{X} \rightarrow O$.

Poznavanjem izvora kovarijata i odabirom ishoda, stvaranje modela nije završeno - potrebno je dodatno odrediti mehanizam predstavljanja znanja i njegove parametre. Jedan od mogućih mehanizama predstavljanja znanja opisan je u nastavku.



Slika 2.2: Primjer Bayesove mreže za domenu koronarne bolesti srca.

2.2 Predstavljanje znanja Bayesovim mrežama

Bayesova mreža (Pearl, 1988) je grafički prikaz distribucije vjerojatnosti nad skupom kovarijata. Sastavljena je iz dva dijela: 1) usmjerene mrežne strukture u obliku usmjerenog acikličkog grafa $\mathcal{G}$, te 2) združene distribucije vjerojatnosti P. Ovi su pojmovi detaljnije opisani u nastavku.

Na slici 2.2 predstavljen je graf Bayesove mreže za domenu koronarne bolesti srca. Svako čvorište grafa predstavlja jednu kovarijatu domene, odnosno svaka je kovarijata X_i predstavljena grafički, čvorištem V_i. Usmjereni lukovi na grafu označavaju neposredno povezana čvorišta, od uzroka k posljedici. Tako se iz strukture grafa na slici 2.2 može iščitati da su *prehrana*, *cigarete* i *stres* neposredni uzroci *KBS*-a, dok su neposredne posljedice (simptomi) *KBS*-a čvorišta *umor* i *EKG*. Također, *stres* je neposredan uzrok *tlaku*, dok je *anemija* neposredan uzrok *umoru*. Kako su ovisnosti predstavljene lukovima, tako su neovisnosti predstavljene nedostatkom lukova; npr. *tlak* nije neposredno vezan uz *umor*, *prehrana* nije neposredno vezana uz *EKG*. No, nedostatak neposredne veze među čvorištima ne znači potpunu neovisnost čvorišta. Posredno povezana čvorišta u grafu uvjetno su neovisna kada su ispunjeni određeni uvjeti (sekcija 4.1.1).

Opisana struktura grafa Bayesove mreže predstavlja kvalitativni opis domene. Kvantitativni dio, onaj koji lokalizirano i količinski određuje odnose među povezanim čvorištima, određen je združenom distribucijom vjerojatnosti, odnosno skupom lokalnih (uvjetnih) distribucija vjerojatnosti (Kjaerulff & Madsen, 2007).

Na slici 2.3 predstavljene su tablice lokalnih (uvjetnih) distribucija vjerojatnosti za pretpostavljenu domenu koronarne bolesti srca, po jedna za svaku kovarijatu, to jest čvorište. Kvantitativno je domena određena združenom distribucijom vjerojatnosti:

Prehrana (V_P)

Dobra (V_{P0})	0,50
Loša (V_{P1})	0,50

Cigarete (V_C)

Da (V_{C0})	0,30
Ne (V_{C1})	0,70

Stres (V_S)

Prisutan (V_{S0})	0,60
Odsutan (V_{S1})	0,40

EKG (V_E)

KBS	Odsutan	Prisutan
Normalan (V_{E0})	0,90	0,05
Abnormalan (V_{E1})	0,10	0,95

Anemija (V_A)

Prisutna (V_{A0})	0,10
Odsutna (V_{A1})	0,90

Tlak (V_T)

Stres	Prisutan	Odsutan
Normalan (V_{T0})	0,20	0,80
Povišen (V_{T1})	0,80	0,20

Umor (V_U)

KBS	Odsutan	Odsutan	Prisutan	Prisutan
Anemija	Prisutna	Odsutna	Prisutna	Odsutna
Prisutan (V_{U0})	0,80	0,05	0,99	0,70
Odsutan (V_{U1})	0,20	0,95	0,01	0,30

KBS (V_K)

Prehrana	Dobra	Dobra	Dobra	Dobra	Loša	Loša	Loša	Loša
Cigarete	Da	Da	Ne	Ne	Da	Da	Ne	Ne
Stres	Prisutan	Odsutan	Prisutan	Odsutan	Prisutan	Odsutan	Prisutan	Odsutan
Odsutan (V_{K0})	0,30	0,70	0,60	0,90	0,01	0,20	0,20	0,60
Prisutan (V_{K1})	0,70	0,30	0,40	0,10	0,99	0,80	0,80	0,40

Slika 2.3: Tablice lokalnih (uvjetnih) distribucija vjerojatnosti za Bayesovu mrežu koronarne bolesti srca.

$$\begin{aligned} \mathrm{P}(\mathcal{G}) &= \mathrm{P}(V_P, V_C, V_S, V_A, V_T, V_E, V_U, V_K) \\ &= \prod_{i=1}^{n} \mathrm{P}(V_i | \boldsymbol{\pi}(V_i)) \\ &= \mathrm{P}(V_P) \cdot \mathrm{P}(V_C) \cdot \mathrm{P}(V_S) \cdot \mathrm{P}(V_A) \cdot \mathrm{P}(V_T | V_S) \\ &\quad \cdot \mathrm{P}(V_E | V_K) \cdot \mathrm{P}(V_U | V_K, V_A) \cdot \mathrm{P}(V_K | V_P, V_C, V_S), \end{aligned} \tag{2.1}$$

gdje je $\boldsymbol{\pi}(V_i)$ skup roditelja čvorišta V_i. Postupak je objašnjen u dodatku B. Izraz 2.1 sadrži informaciju o strukturi mreže, to jest ekvivalentan je grafu sa slike 2.2. Može se koristiti za izračun aposteriorne distribucije vjerojatnosti nekog čvorišta, temeljeno na dokazima (poznatim, to jest opaženim vrijednostima nekog podskupa kovarijata odnosno čvorišta).

Pretpostavimo da želimo odrediti aposteriornu vjerojatnost prisutnosti koronarne bolesti srca za neku osobu, na temelju njoj opaženih vrijednosti svih kovarijata. Za bilo koji skup dokaza, aposteriorna distribucija koronarne bolesti srca određena je izrazom uvjetne vjerojatnosti:

$$\mathrm{P}(V_K | V_P, V_C, V_S, V_A, V_T, V_E, V_U) = \frac{\mathrm{P}(V_P, V_C, V_S, V_A, V_T, V_E, V_U, V_K)}{\mathrm{P}(V_P, V_C, V_S, V_A, V_T, V_E, V_U)}. \tag{2.2}$$

Izraz u brojniku jednak je združenoj distribuciji vjerojatnosti mreže (izraz 2.1), dok se izraz u nazivniku računa marginaliziranjem po izostavljenom čvorištu V_K (zbrajanjem po svim njegovim vrijednostima, što je opisano u dodatku B):

$$\begin{aligned}
&\mathrm{P}(V_P,V_C,V_S,V_A,V_T,V_E,V_U) = \\
&\quad = \sum_K \mathrm{P}(V_P,V_C,V_S,V_A,V_T,V_E,V_U,V_K) \\
&\quad = \mathrm{P}(V_P)\cdot\mathrm{P}(V_C)\cdot\mathrm{P}(V_S)\cdot\mathrm{P}(V_A)\cdot\mathrm{P}(V_T|V_S) \\
&\qquad \cdot\sum_K \mathrm{P}(V_E|V_K)\cdot\mathrm{P}(V_U|V_K,V_A)\cdot\mathrm{P}(V_K|V_P,V_C,V_S).
\end{aligned} \tag{2.3}$$

Konkretan izračun ilustriran je sljedećim primjerom. Pretpostavimo da se neka osoba zdravo hrani (V_{P0}), ali puši (V_{C0}) i izložena je povišenom stresu (V_{S0}) te ima povišen krvni tlak (V_{T1}). Ne boluje od anemije (V_{A1}), ali zamjećuje kronični umor (V_{U0}). Vođena brigom za vlastito zdravlje, odlazi na pretrage te utvrđuje da ima normalnu sliku EKG-a (V_{E0}). Aposteriorna vjerojatnost da ta osoba pati od KBS-a (V_{K1}) je:

$$\begin{aligned}
&\mathrm{P}(V_{K1}|V_{P0},V_{C0},V_{S0},V_{A1},V_{T1},V_{E0},V_{U0}) = \\
&\quad = \frac{\mathrm{P}(V_{P0},V_{C0},V_{S0},V_{A1},V_{T1},V_{E0},V_{U0},V_{K1})}{\mathrm{P}(V_{P0},V_{C0},V_{S0},V_{A1},V_{T1},V_{E0},V_{U0})} \\
&\quad = \Big[\mathrm{P}(V_{P0})\cdot\mathrm{P}(V_{C0})\cdot\mathrm{P}(V_{S0})\cdot\mathrm{P}(V_{A1})\cdot\mathrm{P}(V_{T1}|V_{S0}) \\
&\qquad \cdot\mathrm{P}(V_{E0}|V_{K1})\cdot\mathrm{P}(V_{U0}|V_{K1},V_{A1})\cdot\mathrm{P}(V_{K1}|V_{P0},V_{C0},V_{S0})\Big] \\
&\qquad /\Big[\mathrm{P}(V_{P0})\cdot\mathrm{P}(V_{C0})\cdot\mathrm{P}(V_{S0})\cdot\mathrm{P}(V_{A1})\cdot\mathrm{P}(V_{T1}|V_{S0}) \\
&\qquad \cdot\sum_K \mathrm{P}(V_{E0}|V_K)\cdot\mathrm{P}(V_{U0}|V_K,V_{A1})\cdot\mathrm{P}(V_K|V_{P0},V_{C0},V_{S0})\Big] \\
&\quad = \frac{\mathrm{P}(V_{E0}|V_{K1})\cdot\mathrm{P}(V_{U0}|V_{K1},V_{A1})\cdot\mathrm{P}(V_{K1}|V_{P0},V_{C0},V_{S0})}{\sum_K \mathrm{P}(V_{E0}|V_K)\cdot\mathrm{P}(V_{U0}|V_K,V_{A1})\cdot\mathrm{P}(V_K|V_{P0},V_{C0},V_{S0})} \\
&\quad = \frac{0.05\cdot 0.7\cdot 0.7}{0.9\cdot 0.05\cdot 0.3 + 0.05\cdot 0.7\cdot 0.7} \\
&\quad \simeq 0.64.
\end{aligned} \tag{2.4}$$

Svakom čvorištu pridružen je indeks 0 ili 1, ovisno o tome koju je rednu vrijednost iz tablica lokalnih distribucija vjerojatnosti poprimilo (slika 2.3); npr. V_{K1} označava da je KBS prisutan. Ako bi se za istu osobu utvrdilo da ima loš (abnormalan) EKG (V_{E1}), aposteriorna vjerojatnost KBS-a bila bi:

$$\mathrm{P}(V_{K1}|V_{P0},V_{C0},V_{S0},V_{A1},V_{T1},V_{E1},V_{U0}) = \frac{0.95\cdot 0.7\cdot 0.7}{0.1\cdot 0.05\cdot 0.3 + 0.95\cdot 0.7\cdot 0.7} \simeq 1. \tag{2.5}$$

Ovo je primjer zaključivanja na osnovi potpunog promatranja (engl. *fully-observed case*). Kada bismo htjeli iz zadanog modela odrediti aposteriornu vjerojatnost KBS-a osobe iz prvog scenarija (EKG slika dobra), ali uz nepoznato (nepromotreno) stanje

stresa, tada govorimo o zaključivanju na osnovi nepotpunog promatranja. To se rješava marginaliziranjem po nepoznatom čvorištu:

$$\begin{aligned}
&\mathrm{P}(V_{K1}|V_{P0}, V_{C0}, V_{A1}, V_{T1}, V_{E0}, V_{U0}) = \\
&\quad = \frac{\sum_S \mathrm{P}(V_{P0}, V_{C0}, V_S, V_{A1}, V_{T1}, V_{E0}, V_{U0}, V_{K1})}{\sum_S \mathrm{P}(V_{P0}, V_{C0}, V_S, V_{A1}, V_{T1}, V_{E0}, V_{U0})} \\
&\quad = \Big[\mathrm{P}(V_{E0}|V_{K1}) \cdot \mathrm{P}(V_{U0}|V_{K1}, V_{A1}) \\
&\qquad \cdot \sum_S \mathrm{P}(V_S) \cdot \mathrm{P}(V_{T1}|V_S) \cdot \mathrm{P}(V_{K1}|V_{P0}, V_{C0}, V_S)\Big] \\
&\qquad / \Big[\sum_K \mathrm{P}(V_{E0}|V_K) \cdot \mathrm{P}(V_{U0}|V_K, V_{A1}) \\
&\qquad \cdot \sum_S \mathrm{P}(V_S) \cdot \mathrm{P}(V_{T1}|V_S) \cdot \mathrm{P}(V_K|V_{P0}, V_{C0}, V_S)\Big] \\
&\quad \simeq 0.40.
\end{aligned} \tag{2.6}$$

U nedostatku dokaza o stresu, u obzir dolazi dokaz o povišenom krvnom tlaku. Ukoliko bi tlak bio normalan, vjerojatnost KBS-a bila bi 0.17. Dodatnim uklanjanjem svih preostalih dokaza dobili bismo apriornu vjerojatnost KBS-a. Na isti se način da utvrditi aposteriorna distribucija vjerojatnosti bilo kojeg drugog čvorišta (ili podskupa čvorišta), na osnovi skupa dokaza.

Opisani postupak određivanja aposteriornih vjerojatnosti podskupa čvorišta, temeljeno na dokazima, zove se zaključivanje iz vjerojatnosti (engl. *probabilistic inference*) (Kjaerulff & Madsen, 2007). Slijedom strukture mreže i primjenom tablica uvjetnih distribucija vjerojatnosti, ono se koristi za klasifikaciju nepoznatih primjeraka na osnovi zabilježenih dokaza. Svako je čvorište u Bayesovoj mreži neovisno od svojih ne-potomaka ako su poznate vrijednosti njegovih roditelja (dokaz), stoga je zaključivanje računski brzo (Pearl, 1988). Npr. ako je za Bayesovu mrežu koronarne bolesti srca (slika 2.2) poznato stanje stresa, informacija o tlaku za dijagnozu prisutnosti KBS-a postaje irelevantna (i obratno).

2.3 Analiza preživljenja

Bayesova mreža iz prethodne sekcije primjer je modela predstavljanja znanja. Poznajući vrijednosti nekih kovarijata, tim je modelom moguće predvidjeti distribucije vjerojatnosti nekih drugih varijata. Na primjeru domene koronarne bolesti srca, za poznate vrijednosti kovarijata *prehrana*, *cigarete*, *stres*, *anemija*, *tlak*, *EKG* i *umor*, moguće je odrediti vjerojatnost koronarne bolesti srca. Za Bayesovu mrežu KBS-a može se reći da je klasifikacijski model jer je sposobna klasificirati nove primjerke (dodijeliti ih jednom od razreda).

Za razliku od klasifikacije, analiza preživljenja (Klein & Moeschberger, 2003) bavi se modeliranjem vremena do zbivanja događaja od interesa. Ona nam omogućuje modeli-

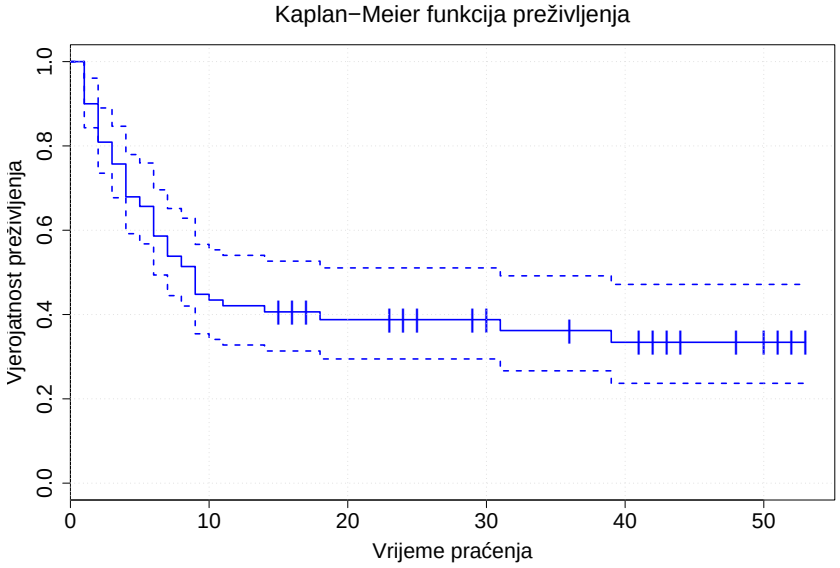
ranje distribucije vjerojatnosti ishoda (klasfikacije) u vremenu. Kako bi to bilo moguće, na postojeći skup kovarijata u domeni potrebno je dodati vremensku dimenziju, tzv. vrijeme praćenja (engl. *observation time*, *follow-up time*). Vrijeme praćenja opisuje u kojem je trenutku za neki primjerak bio zabilježen neki ishod.

Postojećem modelu koronarne bolesti srca (dodatak A) dodana je kovarijata T koja predstavlja vrijeme praćenja. Vrijeme praćenja bilježi trenutak nastupa KBS-a (npr. primjećivanje boli u prsima kao blaži simptom ili pak nastup infarkta miokarda, odnosno srčanog udara, kao teži simptom) ili prestanka praćenja stanja pacijenta iz drugih razloga (KBS odsutan). Logično je očekivati korelaciju između odabranih životnih navika (prehrana, pušenje, stres) i vremena praćenja primjeraka s ustanovljenim KBS-om. U tom je slučaju vrijeme praćenja jednako vremenu preživljenja (pojašnjeno u idućoj sekciji). Npr. za osobu koja se nekvalitetno hrani, puši i pod stalnim je stresom, postoji velika vjerojatnost da će prije 40-e godine života oboljeti od KBS-a, za razliku od osobe koja se zdravo hrani, ne puši i nije pod stresom. Pored toga, velika je vjerojatnost da ta druga osoba nikada neće oboljeti od KBS-a.

Za razliku od vremena praćenja, vrijeme preživljenja (engl. *survival time*) može se definirati kao vrijeme do zbivanja događaja od interesa (engl. *time-to-event occurrence*). Taj događaj može biti npr. razvoj neke bolesti, reakcija na tretman, ponavljanje bolesti ili smrt (Lee & Wang, 2003). Vrijeme preživljenja je vrijeme proteklo od početka promatranja do trenutka zbivanja ishoda, u ovom primjeru vrijeme do utvrđivanja prisutnosti KBS-a. Primjerci s niskom vjerojatnošću prisutnosti KBS-a teoretski bi trebali imati beskonačno dugo vrijeme preživljenja. No, kao i u ostalim biološkim organizmima i mehaničkim sustavima, u kojima sve ima svoj vijek trajanja, vrijeme preživljenja ne može biti beskonačno (npr. zbog završetka studije ili smrti primjerka iz drugih razloga). Ukoliko je vrijeme praćenja nekog primjerka kraće od vremena preživljenja, kaže se da je on cenzuriran (npr. KBS nekoj osobi nije bio ustanovljen do kraja studije, ili je ta osoba stradala u prometnoj nesreći, ili je pak napustila studiju zbog drugih razloga, npr. preseljenja u drugi grad). Implikacije cenzure detaljnije su razjašnjene u sekciji 2.3.2.

Pretpostavimo da je svaki primjerak $\mathbf{x}$ opisan m-dimenzionalnim vektorom kovarijata $\mathbf{X} = (X_1, \ldots, X_m)$. Vjerojatnost preživljenja do trenutka t za neki primjerak $\mathbf{x}$ određena je izrazom $S(t) = \mathrm{P}(T > t)$, u kojem je T pozitivna slučajna varijata koja predstavlja vrijeme preživljenja primjerka $\mathbf{x}$. Modeliranje preživljenja znači nalaženje odnosa među vrijednostima skupa kovarijata $\mathbf{X}$ i funkcije preživljenja $S(t)$.

Primjer funkcije preživljenja za domenu koronarne bolesti srca ilustriran je na slici 2.4. Korišteni podaci predstavljaju pretpostavljeni uzorak od 100 osoba u svojoj 40-oj godini života. Vrijeme praćenja izraženo je u godinama. Svim osobama kojima je zabilježen KBS, vrijeme praćenja predstavlja vrijeme preživljenja; ostalima predstavlja vrijeme stvarne smrti iz drugih razloga. Zanemarivanjem vremena praćenja ovaj problem ponovno postaje klasifikacijski, jer stvarne cenzure zapravo nema. Svaki pad veličine funkcije preživljenja predstavlja neki broj zabilježenih ishoda u tom trenutku (ustanovljen KBS). Na mjestima na krivulji obilježenim vertikalnom crticom, zabilježena je jedna ili više smrti iz drugih razloga nevezanih uz KBS. Iz krivulje je vidljivo



Slika 2.4: Funkcija preživljenja $S(t)$ za domenu koronarne bolesti srca. Funkcija je praćena krivuljama 95%-tnih intervala pouzdanosti (crtkano).

kako je najveći udio zabilježenih oboljenja od KBS-a prisutan upravo između 40-e i 50-e godine života osoba iz uzorka. Interval pouzdanosti se s vremenom širi zbog stalnog smanjenja veličine uzorka u vremenu. Detalji o korištenim podacima za izračun funkcije preživljenja sa slike 2.4 mogu se naći u dodatku A.

Funkcija je modelirana jednim od neparametarskih pristupa, tzv. procjenom Kaplan - Meier (engl. *Kaplan - Meier product limit estimate*) (Kaplan & Meier, 1958). Procjena funkcije preživljenja određena je izrazom:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{N_i - d_i}{N_i} \right) = \hat{S}(t-1)\left(1 - \frac{d_t}{N_t}\right), \tag{2.7}$$

u kojem je d_i broj događaja zabilježenih u trenutku t_i, u kojem se zbio jedan ili više događaja, dok je N_i broj još promatranih primjeraka u trenutku t_i. Procjena Kaplan - Meier pretpostavlja da je cenzuriranje slučajno, to jest da su vremena cenzuriranja neovisna od vremena preživljenja. Iz tog razloga nije primjerena u slučajevima u kojima je primjer cenzuriran iz razloga povezanih s uzrocima zbivanja događaja (Lee & Wang, 2003). Funkcije preživljenja monotono su padajuće; veći pad funkcije u nekom području označava veći broj zabilježenih ispada.

Funkcije, blisko povezane s funkcijom preživljenja, su gustoća vjerojatnosti i hazard (Lee & Wang, 2003). Funkcija gustoće vjerojatnosti $f(t)$ (engl. *probability density function*) određena je izrazom:

$$f(t) = -\frac{d}{dt}S(t). \tag{2.8}$$

Iz nje izvedena funkcija kumulativne distribucije $F(t)$ određena je izrazom:

$$F(t) = \int_0^t f(x)dx = 1 - S(t). \tag{2.9}$$

Funkcija hazarda $h(t)$ (engl. *hazard function*) daje vjerojatnost ispada (zbivanja događaja) u vrlo kratkom vremenskom intervalu:

$$h(t) = -\frac{f(t)}{S(t)} = -\frac{d}{dt}\log S(t). \tag{2.10}$$

Ona daje trenutni potencijal ispada primjerka u trenutku t za svaki primjerak koji je praćen do tog trenutka. Iz nje izvedena funkcija kumulativnog hazarda $H(t)$ određena je izrazom:

$$H(t) = \int_0^t h(x)dx = -\log S(t). \tag{2.11}$$

2.3.1 Postupci modeliranja preživljenja

Svrha je prognostičkih modela izvesti generalizirana pravila iz nekih podataka koja bi se dala primijeniti za predviđanje budućih ishoda novih primjeraka, koji nisu bili uključeni u uzorku za učenje. Generalizacija se obično dobiva dodjeljivanjem različitih kombinacija vrijednosti kovarijata različitim skupovima primjeraka, u analizi preživljenja obično zvanih rizične skupine (engl./lat. *risk strata*), označenih s $\mathscr{X}_1, \ldots, \mathscr{X}_s$. Svaka rizična skupina $\mathscr{X}_i$ sadrži više primjeraka sličnih karakteristika i može biti predstavljena jednom jedinom procjenom funkcije preživljenja $\hat{S}_i(t)$. Pošto je u stvarnim domenama veličina uzorka obično ograničena, nikada nisu zastupljene sve kombinacije prostora vrijednosti kovarijata $\mathbf{X}$.

Postoji pregršt postupaka za modeliranje preživljenja. Jedan od njih predstavljen je u nastavku. Prognostički indeks koronarne bolesti srca određen je izrazom:

$$PI(\mathbf{X}) = \boldsymbol{\beta} \cdot \mathbf{X'} = \beta_P{\cdot}X_P + \beta_C{\cdot}X_C + \beta_S{\cdot}X_S + \beta_T{\cdot}X_T + \beta_A{\cdot}X_A + \beta_E{\cdot}X_E + \beta_U{\cdot}X_U, \tag{2.12}$$

gdje je $\boldsymbol{\beta}$ skup regresijskih koeficijenata. Prognostički indeks, poznat pod nazivom omjer hazarda (engl. *hazard ratio*), određen je prirodnim logaritmom kvocijenta hazarda i osnovnog hazarda. Iz toga slijedi funkcija preživljenja:

$$S(t|\mathbf{X}) = e^{-H_0(t)\cdot PI(\mathbf{X})}, \tag{2.13}$$

u kojoj je $H_0(t)$ funkcija kumulativnog osnovnog hazarda (engl. *cumulative baseline hazard*). Ovaj se model predstavljanja znanja u analizi preživljenja naziva regresijski

Tablica 2.1: Modeli proporcionalnih hazarda za domenu koronarne bolesti srca. β je regresijski koeficijent, SE standardna pogreška, z kvocijent β i SE (Waldova statistika), a p empirijska razina statističke značajnosti.

(a) Uključene sve kovarijate

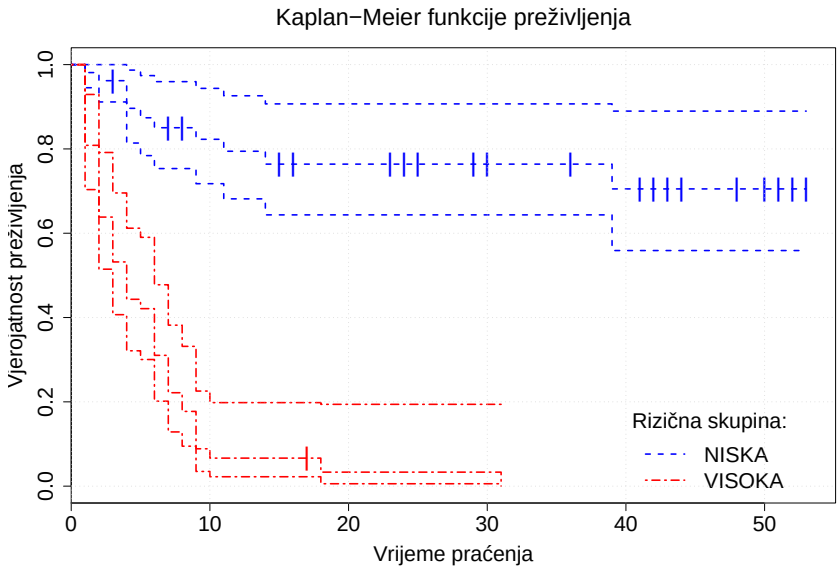
Kovarijata	β	e^{β}	SE	z	p
Prehrana	0.1352	1.145	0.302	0.448	0.65000
Cigarete	-0.6749	0.509	0.316	-2.133	0.03300
Stres	-0.6227	0.536	0.386	-1.615	0.11000
Tlak	-0.0702	0.932	0.320	-0.220	0.83000
Anemija	0.4548	1.576	0.535	0.850	0.40000
EKG	2.8764	17.750	0.758	3.796	0.00015
Umor	-1.2766	0.279	0.376	-3.399	0.00068

(b) Uključene sve kovarijate osim *EKG*-a

Kovarijata	β	e^{β}	SE	z	p
Prehrana	0.330	1.392	0.301	1.098	0.27000
Cigarete	-0.503	0.605	0.310	-1.622	0.10000
Stres	-0.670	0.512	0.375	-1.789	0.07400
Tlak	0.276	1.317	0.311	0.886	0.38000
Anemija	0.720	2.055	0.528	1.366	0.17000
Umor	-2.046	0.129	0.381	-5.371	0.00000

model proporcionalnih hazarda (Cox, 1972). Postupak je detaljnije opisan u sekciji 4.4.2.

Tablica 2.1 sadrži procijenjene regresijske koeficijente domene koronarne bolesti srca (dodatak A) za dva slučaja. Koeficijenti su izračunati korištenjem alata R (sekcija 1.2). U prvom su slučaju pokrivene sve kovarijate, dok drugi pretpostavlja sve kovarijate osim *EKG*-a. U drugom je slučaju ta kovarijata namjerno izostavljena jer je previše korelirana s ishodom, što daje gotovo savršen prediktor; upravo će se taj slabiji model koristiti u nastavku za prikaz primjera klasifikacije. Regresijski model proporcionalnih hazarda u oba se slučaja može tumačiti razmatranjem procijenjenih regresijskih koeficijenata pojedinačno: pozitivan koeficijent povećava hazard, dok ga negativan smanjuje, u iznosu veličine njegove apsolutne vrijednosti. Tako npr. odsutnost *anemije* povećava rizik *KBS*-a, dok ga odsutnost *stresa* smanjuje. Kao statistički značajna za ishod, s empirijskom razinom $p < 0.05$, u drugom je slučaju odabrana jedino kovarijata *umor* (u prvom slučaju su to, pored *umora*, još kovarijate *EKG* i *cigarete*). Empirijska razina statističke značajnosti računa se dvostranim testom površine ispod grafa normalne distribucije prema Waldovoj statistici z, koja je asimptotski standardno normalna pod pretpostavkom da je njoj pripadajući regresijski koeficijent β jednak nuli.



Slika 2.5: Predviđanje regresijskog modela proporcionalnih hazarda za domenu koronarne bolesti srca. Funkcije preživljenja praćene su krivuljama 95%-tnih intervala pouzdanosti.

Podjela primjeraka na rizične skupine ilustrirana je na modelu iz drugog slučaja tablice 2.1 (bez *EKG*-a). Za podjelu je korišten prilagođen prognostički indeks $PI(\mathbf{X}) = \boldsymbol{\beta} \cdot \mathbf{X'}$ kao linearni prediktor: ako je $PI(\mathbf{x}) \geq 0$, tada svrstaj $\mathbf{x}$ u visokorizičnu skupinu (KBS prisutan); u protivnom svrstaj $\mathbf{x}$ u niskorizičnu skupinu (KBS odsutan). Postojeći primjerci, koji su bili dodijeljeni jednoj ili drugoj rizičnoj skupini $\mathscr{X}_i$, mogu se koristiti pri određivanju procjene funkcija preživljenja $\hat{S}_i(t)$ svake skupine ponaosob. Dobivene funkcije preživljenja predstavljene su na slici 2.5. Krivulja preživljenja koja opisuje niskorizičnu skupinu sporije pada od one koja opisuje visokorizičnu skupinu; to bi značilo da visokorizična skupina bilježi puno više ispada (utvrđenih KBS-ova) od one niskorizične. Iako se već na prvi pogled iz ove slike može procijeniti modelom utvrđena kakvoća razdvajanja rizičnih skupina (koja je na ovom primjeru dobra), ponekad ta okvirna procjena zna zavarati. Kvalitetnije je rješenje za procjenu kakvoće razdvajanja rizičnih skupina log-rank test (Lee & Wang, 2003; Peto & Peto, 1972).

Log-rank test statistički je test nulte hipoteze, koja pretpostavlja da ne postoji značajna razlika između distribucija preživljenja dvaju uzoraka. Učestalo se koristi u kliničkim studijama radi utvrđivanja učinkovitosti novih lijekova na nekoj skupini pacijenata, u odnosu na kontrolnu skupinu (uglavnom placebo). Računa se izrazom:

$$\chi^2 \approx \frac{D^A - E^A}{E^A} + \frac{D^B - E^B}{E^B}, \tag{2.14}$$

u kojem su D^A i D^B ukupni brojevi zabilježenih događaja u skupini A odnosno B, dok

su E^A i E^B brojevi očekivanih događaja, koji se računaju izrazima:

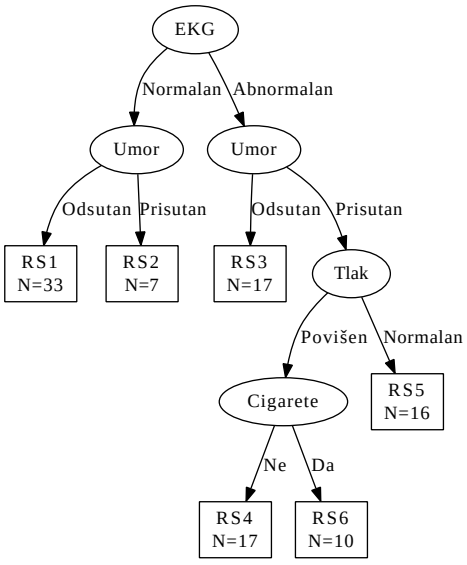
$$E^A = \sum_{i:t_i} \frac{N_i^A}{N_i^A + N_i^B} \cdot d_i, \tag{2.15}$$

$$E^B = \sum_{i:t_i} \frac{N_i^B}{N_i^A + N_i^B} \cdot d_i, \tag{2.16}$$

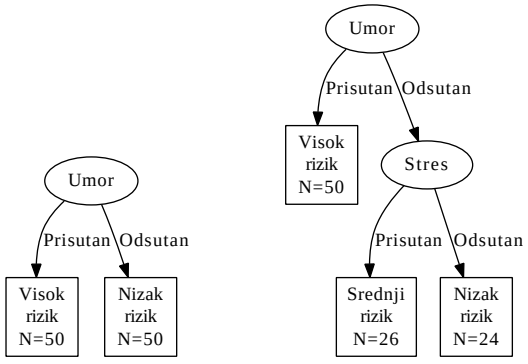
u kojima je d_i broj događaja zabilježenih u trenutku t_i, gdje se zbio jedan ili više događaja, dok je N_i^A broj još promatranih primjeraka u trenutku t_i u skupini A, odnosno N_i^B je broj još promatranih primjeraka u trenutku t_i u skupini B. Testna statistika približno je jednaka hi-kvadratnoj distribuciji s jednim stupnjem slobode; visoka vrijednost $\mathcal{X}^2$ značila bi odbacivanje nulte hipoteze, odnosno utvrdila bi značajnu razliku među skupinama uz unaprijed određenu razinu pouzdanosti (Lee & Wang, 2003). Za rizične skupine sa slike 2.5 log-rank statistika iznosi $\mathcal{X}^2 = 64.0374$, što znači da je nulta hipoteza odbačena na statističkoj razini značajnosti testa $\alpha = 0.01$ uz 1 stupanj slobode (broj rizičnih skupina umanjen za jedan), odnosno da su rizične skupine različite. Podaci korišteni za učenje modela upotrijebljeni su i za izračun log-rank statistike (u ovom poglavlju radi jednostavnosti nisu korišteni posebni podaci za testiranje modela).

Pored opisanog regresijskog modela proporcionalnih hazarda, postoji još jedan popularan način modeliranja preživljenja - klasifikacijska i regresijska stabla (engl. *classification and regression trees*, CART) (Breiman *et al.*, 1984). Stabla se grade rekurzivnim razdjeljivanjem skupa podataka na način *zavadi pa vladaj* (lat. *divide et impera*). U svakom se koraku odabere i doda u stablo ona kovarijata, koja najbolje razdvaja skup podataka. Kvaliteta razdvajanja kriterijska je funkcija, mjerena homogenušću primjeraka unutar čvorišta i heterogenošću primjeraka među čvorištima. U analizi preživljenja ta se veličina (kriterijska funkcija) najčešće računa korištenjem opisane log-rank statistike. Građenje stabla završava onog trenutka kada novim razdjeljivanjem podataka nije moguće povisiti vrijednost kriterijske funkcije. CART je u velikoj mjeri sličan ID3 algoritmu (Quinlan, 1986) - razlikuju se jedino u terminologiji i mjerama rekurzivnog razdjeljivanja.

Na slici 2.6 prikazan je model CART za domenu koronarne bolesti srca (dodatak A). Model uključuje kovarijate *EKG*, *umor*, *tlak* i *cigarete*. U svakom od šest listova stabla, koja predstavljaju pojedine rizične skupine, naveden je broj primjeraka koji spadaju u tu skupinu (iz uzorka od 100 primjeraka). S obzirom na to da je broj rizičnih skupina koje predlaže model prevelik, njihov je broj potrebno smanjiti kako bi model bio upotrebljiv za možebitnu klasifikaciju. To se jednostavno da učiniti podrezivanjem stabla (engl. *pruning*). Na slici 2.7 predstavljena su dva modela CART, naučena iz svih kovarijata osim *EKG*-a, kako bi primjer odgovarao onom primijenjenom na regresijskom modelu proporcionalnih hazarda. Stabla su podrezana uz parametre kompleksnosti (engl. *complexity parameter*) $cp = 0.1$ i $cp = 0.05$, te predlažu podjelu na dvije odnosno tri rizične skupine. Parametar kompleksnosti cp, odnosno regularizacijski parametar (engl. *regularization parameter*), određuje kompromis između ukupne rezidualne sume kvadrata pogrešaka i složenosti modela uzimajući u obzir broj listova stabla (Bishop,



Slika 2.6: Model CART za domenu koronarne bolesti srca. U model su uključene sve kovarijate, bez podrezivanja.



(a) $cp = 0.1$ (b) $cp = 0.05$

Slika 2.7: Modeli CART za domenu koronarne bolesti srca. Oba modela naučena su iz svih podataka, osim za kovarijatu *EKG*, koja je isključena. Modeli su naknadno podrezani za različite parametre kompleksnosti *cp*.

2007). Kao i u modelu, predstavljenom u tablici 2.1(b), kovarijata *umor* prepoznata je kao najutjecajnija; slijedi joj također bitna kovarijata *stres* na modelu s tri rizične skupine na slici 2.7(b). Taj model sve primjerke kod kojih je prisutan *umor* svrstava u visokorizičnu skupinu; uz odsutan *umor*, ali prisutan *stres*, svrstava ih u srednjerizičnu skupinu - u protivnom ih svrstava u niskorizičnu skupinu. Funkcije preživljenja za modele sa slike 2.7 predstavljene su na slici 2.8.

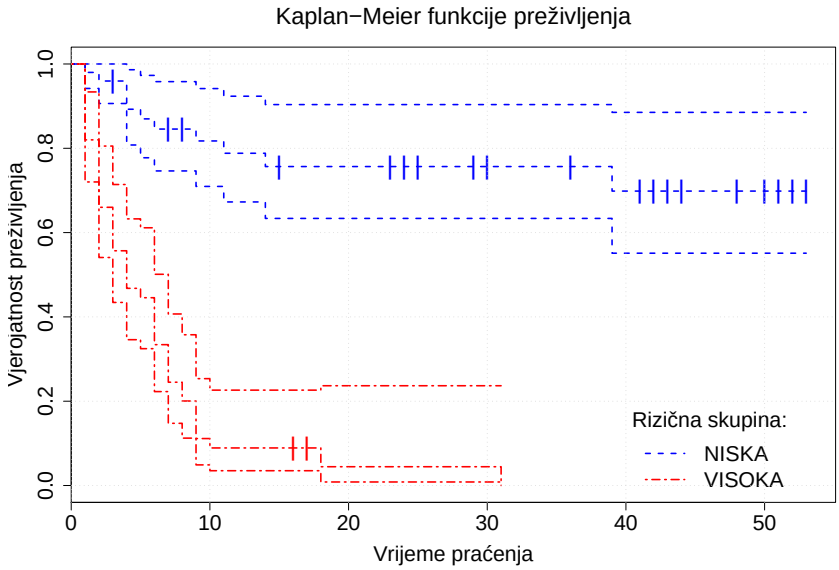
2.3.2 Cenzura

Važna posebnost podataka o preživljenju, koja stvara razliku između analize preživljenja i standardne klasifikacije, jest cenzuriranje. Općenito, do cenzure dolazi kada događaj od interesa nije bio opažen za neki primjerak. U nekoj kliničkoj studiji npr. cenzura bi podrazumijevala preuranjeno povlačenje pacijenta iz studije, iz razloga koji nisu povezani s opažanim događajem od interesa. Npr. cenzuriranje u studiji koronarne bolesti srca dogodi se ako je promatrana osoba podlegla ozljedama zadobivenim u prometnoj nezgodi, ili se odselila u drugu državu za vrijeme trajanja studije, prije nego što je događaj od interesa (ustanovljen KBS) mogao biti zabilježen.

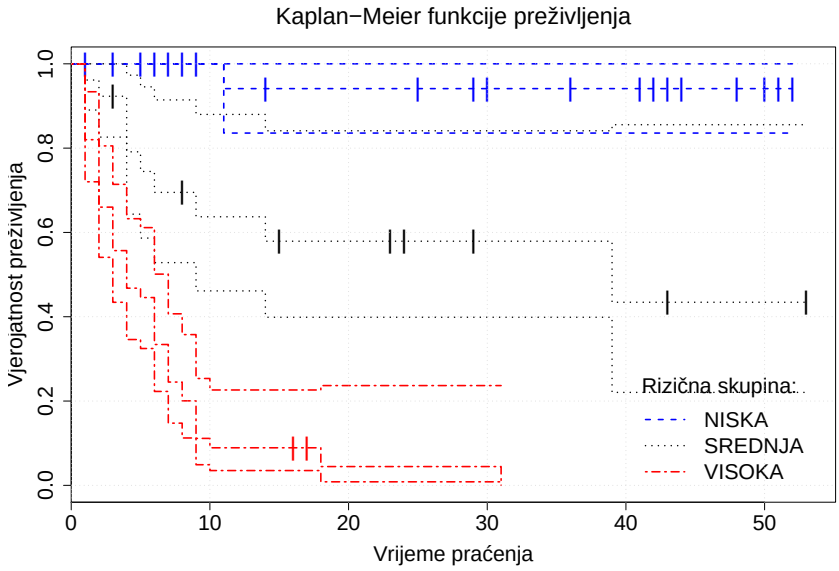
Kleinbaum (2005) daje tri moguća uzroka pojave cenzuriranja: (1) osoba ne doživi događaj od interesa za vrijeme trajanja studije; (2) osoba je izgubljena iz praćenja; (3) osoba je povučena iz studije zbog smrti, nevezane uz promatrani događaj. Ovi se oblici cenzuriranja često zovu desno cenzuriranje. Na slici 2.9 predstavljen je primjer praćenja statusa osoba u pretpostavljenoj studiji koronarne bolesti srca. Početak studije označen je vremenskom oznakom t_P, dok je kraj označen sa t_K. Predstavljeno je praćenje šest osoba u tom intervalu sa stvarnim događajima povezanim s prije opisanim mogućim uzrocima pojave cenzuriranja. Pretpostavimo da se osoba $\mathbf{x}_A$ za vrijeme trajanja studije preselila u drugu državu. Ishod za nju nije zabilježen, dakle radi se o cenzuri zbog uzroka (1). Za osobe $\mathbf{x}_B$ i $\mathbf{x}_C$ u vremenu trajanja studije nije bio zabilježen KBS, dakle radi se o cenzuri zbog uzroka (2). Osoba $\mathbf{x}_D$ u nekom je trenutku smrtno stradala u prometnoj nesreći - to je cenzura zbog uzroka (3). Za osobe $\mathbf{x}_E$ i $\mathbf{x}_F$ je u vremenu praćenja ustanovljen KBS, to jest jedino one nisu cenzurirane.

Na slici 2.10 predstavljena je funkcija preživljenja za domenu koronarne bolesti srca nakon umjetnog 60%-tnog cenzuriranja. To znači da je približno 60% primjeraka s ustanovljenim KBS-om pretvoreno u cenzurirane primjerke, to jest bilo je pridruženo onoj skupini primjeraka za koje KBS nije bio ustanovljen. Kod podataka o preživljenju u stvarnom svijetu veoma je teško (ponekad i nemoguće) razlikovati te dvije skupine cenzuriranih podataka - onih za koje događaj neće biti zabilježen (niti po završetku praćenja), te onih za koje bi događaj od interesa bio zabilježen, da su bili dovoljno dugo praćeni. U usporedbi s funkcijom preživljenja izvornog (necenzuriranog) uzorka na slici 2.4, ova sporije pada iz očitog razloga: 60% ispada ovdje nije zabilježeno zbog cenzure.

Cenzuriranje u podacima o preživljenju glavni je razlog zbog kojeg se standardne metode (nadziranog) strojnog učenja obično ne koriste za učenje modela preživljenja. Konačan opaženi ishod za neki primjerak može biti ili zabilježen, ili cenzuriran (stoga djelomično nepoznat). Čak i ako nas zanima samo modeliranje konačnog ishoda, bez obzira na vrijeme, cenzurirani podaci o preživljenju predstavljaju značajan problem

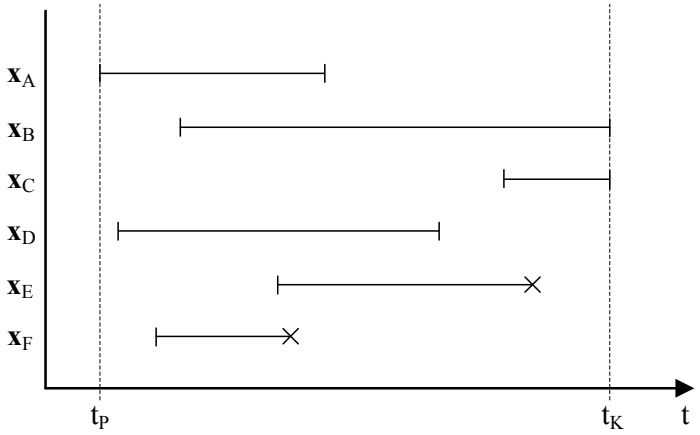


(a) Predviđanje za model sa slike 2.7(a)

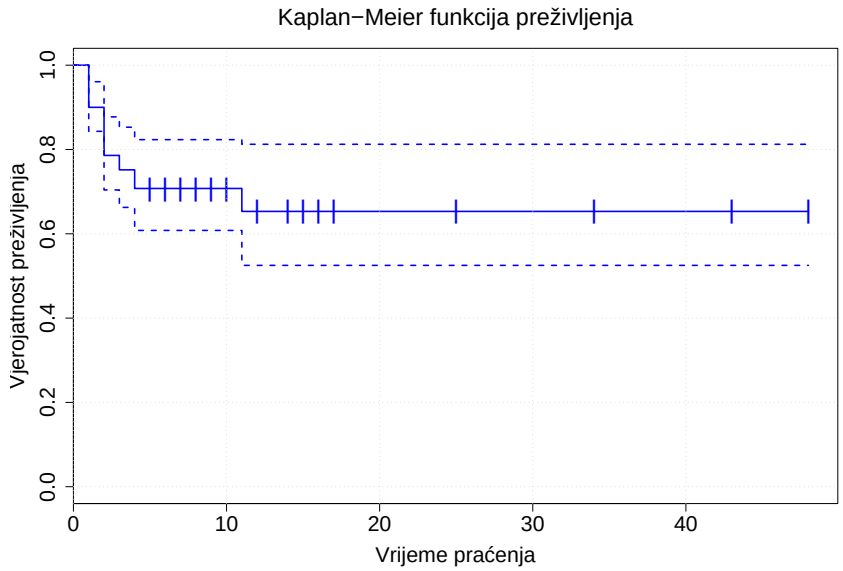


(b) Predviđanje za model sa slike 2.7(b)

Slika 2.8: Predviđanja podrezanih modela CART za domenu koronarne bolesti srca. Funkcije preživljenja praćene su krivuljama 95%-tnih intervala pouzdanosti.



Slika 2.9: Primjer praćenja primjeraka u uzorku u vremenskom intervalu (t_P, t_K), uz prisutnost cenzure. Na kraju praćenja su križićem označeni primjerci za koje se događaj od interesa zbio ($\mathbf{x}_E$ i $\mathbf{x}_F$), dok su vertikalnom crticom označeni cenzurirani primjerci ($\mathbf{x}_A$, $\mathbf{x}_B$, $\mathbf{x}_C$ i $\mathbf{x}_D$).



Slika 2.10: Funkcija preživljenja $S(t)$ za domenu koronarne bolesti srca, nakon uvođenja približno 60%-tne umjetne cenzure u uzorku. Funkcija je praćena krivuljama 95%-tnih intervala pouzdanosti (crtkano).

Tablica 2.2: Log-rank statistika $\mathcal{X}^2$ kakvoća razdvajanja rizičnih skupina modela naučenih iz domene koronarne bolesti srca (bez kovarijate *EKG*), sa i bez cenzure. Testirani modeli su: regresijski model proporcionalnih hazarda (PH), model CART s dva (CART--2) i tri (CART-3) lista, te naivni Bayesov klasifikator (NB).

	Bez cenzure		60%-tna cenzura	
Model	$\mathcal{X}^2$	p	$\mathcal{X}^2$	p
PH	64.0374	$1.2212 \cdot 10^{-15}$	20.4235	$6.2062 \cdot 10^{-6}$
CART-2	56.0566	$7.0388 \cdot 10^{-14}$	19.2220	$1.1637 \cdot 10^{-5}$
CART-3	59.8784	$1.0103 \cdot 10^{-14}$	19.2220	$1.1637 \cdot 10^{-5}$
NB	50.7016	$1.0754 \cdot 10^{-12}$	1.5005	0.22059261

standardnim postupcima strojnog učenja. Cenzuriranje u podacima može se u nekoj mjeri smatrati šumom u ishodu.

Tablica 2.2 prikazuje kakvoće razdvajanja rizičnih skupina modela naučenih iz domene koronarne bolesti srca, s i bez cenzure (dodatak A). Pored do sada predstavljenih modela preživljenja (tablica 2.1/b/ i slika 2.7), u tablici se nalaze i vrijednosti za naivni Bayesov klasifikator. Naivni Bayesov klasifikator, kao predstavnik metoda strojnog učenja, jednostavan je probabilistički klasifikator, temeljen na Bayesovom teoremu i snažnim pretpostavkama neovisnosti. Detaljnije je opisan u sekciji 4.4.1. U odnosu na predstavljene modele preživljenja, naivni Bayesov klasifikator je u stratifikaciji (podjeli na rizične skupine) lošiji, s i bez cenzure. Uz cenzuru postaje čak u potpunosti neupotrebljiv ($\mathcal{X}^2 = 1.5$, $p = 0.22$).

Ako je broj cenzuriranih primjeraka u podacima relativno nizak, u odnosu na njihovu veličinu, njima se u strojnom učenju može rukovati kao da nisu cenzurirani, bez bojazni da će šum u informaciji o razredu značajnije utjecati na uspješnost klasifikatora. Ukoliko je pak taj broj relativno visok, šum u razredu će biti toliko snažan da će veliku većinu modela strojnog učenja učiniti neupotrebljivom. Prethodnom pripremnom obradom cenzuriranih podataka o preživljenju ili prikladnom intervencijom u mehanizme algoritama strojnog učenja, moguće je povećati učinkovitost naučenih modela. Upravo je to tema ove disertacije, s naglaskom na primjenu na učenje Bayesovih mreža.

2.4 Srodni radovi

Ideja korištenja postupaka strojnog učenja u analizi preživljenja stara je 10-ak godina. Nekoliko je popularnih postojećih postupaka strojnog učenja prilagodbom podataka o preživljenju uspješno primijenjeno u stvaranju ekspertnih modela za razne domene kliničke medicine (Biganzoli *et al.*, 1998; Burke *et al.*, 1997; Delen *et al.*, 2005; Ripley & Ripley, 2001; Snow *et al.*, 1994; Zupan *et al.*, 2000). Umjetne neuronske mreže, stabla odlučivanja, naivni Bayesovi klasifikatori i logistički modeli najčešće su korišteni

za opisivanje preživljenja kao funkcije vremena ili konačnog ishoda preživljenja. U novije je vrijeme sve popularnija njihova primjena u analizi genskih zapisa (Evers & Messow, 2008; Hothorn *et al.*, 2006; Kronek & Reddy, 2008). Isto tako, evidentni su i pokušaji korištenja evolucijskih postupaka za učenje prognostičkih modela iz podataka o preživljenju (Peña-Reyes & Sipper, 2000).

Za razliku od ostalih postupaka strojnog učenja, učenje Bayesovih mreža iz cenzuriranih podataka o preživljenju gotovo je u potpunosti zanemareno (Lucas *et al.*, 2004). U nastavku su detaljnije opisani neki pokušaji. Blanco *et al.* (2005) proučavali su izbor kovarijata u Bayesovim klasifikatorima. Kao primjer, koristili su podatke o preživljenju pacijenata oboljelih od ciroze jetre. Podatke su podijelili prema tome je li pacijent umro u prvih 6 mjeseci nakon namještanja TIPS-a ili ne. Radi izbjegavanja pristranosti, cenzurirane primjerke su odstranili iz skupa podataka. Modeli koje su koristili srodni su Bayesovim mrežama: naivni Bayesov klasifikator (s i bez izbora kovarijata), naivni Bayesov klasifikator proširen stablom (engl. *tree-augmented*), polunaivni Bayesov klasifikator (engl. *semi-naive*) i k-ovisni Bayesov klasifikator (engl. *k-dependence*). Sierra & Larranaga (1998) proučavali su učenje Bayesovih mreža korištenjem genetskih algoritama. Postupak su ilustrirali na podacima o malignom melanomu kože. Podatke su podijelili na tri neovisna skupa, ovisno o vremenu praćenja pojedinih primjeraka (preživljenje nakon prve, treće i pete godine). Iz svakog je skupa raznim varijacijama odabranog pristupa naučen zaseban model. Struktura modela naučenih različitim varijacijama odabranog postupka u raznim vremenskim okvirima uspoređena je korištenjem Hammingove udaljenosti (Hamming, 1950). Pristup predstavljen u Marshall *et al.* (2000) rukuje vremenskom dimenzijom u podacima o preživljenju putem dinamičkih Bayesovih mreža (Murphy, 2002). Kombiniranjem Bayesovih mreža i latentnog Markovljevog modela uspješno su spojili uzročno predstavljanje i modeliranje preživljenja. Taj su postupak ilustrirali na podacima o praćenju razdoblja boravka gerijatrijskih pacijenata u bolnici (po fazama).

Gotovo sve navedene radove koji opisuju razne pristupe primjene postupaka strojnog učenja u analizi preživljenja, vežu činjenice: 1) da nisu kvalitetno vrednovani u odnosu na standardne postupke analize preživljenja i 2) da nisu kvalitetno uspoređeni međusobno. Većina radova koji predlažu novi pristup ili primjenjuju postojeći, testiraju ga na jednom do dva skupa podataka, često novih (nestandardnih). Najčešće su to studije slučaja za razna područja kliničke medicine i biomedicine. Testiranje pristupa na umjetno generiranim podacima je rijetkost, kao i statističke usporedbe dobivenih rezultata. Iz navedenih je razloga jedan od ciljeva ove disertacije detaljno predstaviti i temeljito empirijski vrednovati različite popularne pristupe (uz jedan nov) rukovanju cenzuriranim podacima o preživljenju u strojnom učenju. Naglasak je pritom stavljen na učenje Bayesovih mreža u analizi preživljenja, jer je upravo to područje najslabije istraženo. Predložen postupak prilagodbe podataka odstranjivanjem šuma cenzure (sekcija 3.5) razlikuje se od ostalih postupaka prilagodbe podataka o preživljenju po tome što: 1) koristi informaciju o vremenu praćenja isključivo za određivanje apriorne distribucije vjerojatnosti ishoda te 2) mijenja informaciju o ishodu u podacima prije učenja. Ukoliko bi vremena praćenja pojedinih primjeraka bila nepoznata, ali bi istovremeno postojala

ekspertna procjena distribucije vjerojatnosti ishoda na uzorku, taj bi se postupak jedini mogao uspješno primijeniti (ne uzimajući u obzir naivni pristup tretiranju cenzuriranih primjeraka kao negativnih). Predloženi postupak najbolje se nosi s visokocenzuriranim podacima o preživljenju.

Poglavlje 3

Postupci prilagodbe podataka o preživljenju za algoritme strojnog učenja

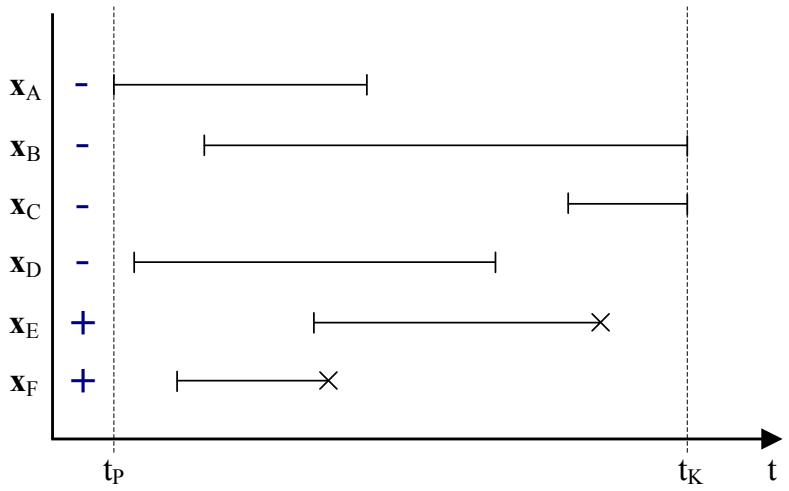
U nastavku su predstavljena neka od rješenja problema strojnog učenja iz cenzuriranih podataka o preživljenju u obliku postupaka prilagodbe takvih podataka. Svim je predstavljenim postupcima zajedničko to, da u nekoj mjeri koriste informaciju o vremenu praćenja te se ta informacija, nakon pripremne obrade, iz podataka u potpunosti briše. Ovo se odnosi na podatke za učenje; u podacima za testiranje se informacija zadržava zbog mogućnosti vrednovanja naučenih modela metrikama korištenim u analizi preživljenja.

Razina cenzure u podacima u ovoj je disertaciji opisana postotkom pozitivnih primjeraka za koje nije bio zabilježen ishod (cenzurirani) u odnosu na ukupan broj pozitivnih primjeraka. Iz toga slijedi gruba podjela na nisku (do $\approx 20\%$), srednju (do $\approx 50\%$) i visoku (do $\approx 80\%$) razinu cenzure, odnosno lakše, srednje i teško cenzurirane podatke (Royston & Sauerbrei, 2004; Štajduhar *et al.*, 2009). To je potrebno naglasiti jer neki autori pod pojmom razina cenzure podrazumijevaju udio svih cenzuriranih primjeraka (stvarno i prividno negativnih) u ukupnom broju primjeraka, što nije pogodno za objašnjenja nekih koncepata predstavljenih u nastavku.

3.1 Tretiranje cenzuriranih primjeraka kao negativnih

Naivni pristup rukovanju cenzuriranim podacima o preživljenju jest tretiranje cenzuriranih primjeraka kao negativnih. Bez obzira na vrijeme praćenja, svaki se cenzurirani primjerak tretira kao taj, za koji je ishod negativan (slika 3.1), to jest vrijeme praćenja se u potpunosti zanemaruje.

Naivni pristup može biti iznimno loš u slučajevima u kojima je primjerak cenzuriran nakon kratkog praćenja, iz razloga što se teško može procijeniti konačni ishod za dani primjerak (vjerojatnost preživljenja približno je jednaka apriornoj vjerojatnosti pre-



Slika 3.1: Primjer tretiranja cenzuriranih primjeraka kao negativnih ($\mathbf{x}_A$, $\mathbf{x}_B$, $\mathbf{x}_C$ i $\mathbf{x}_D$).

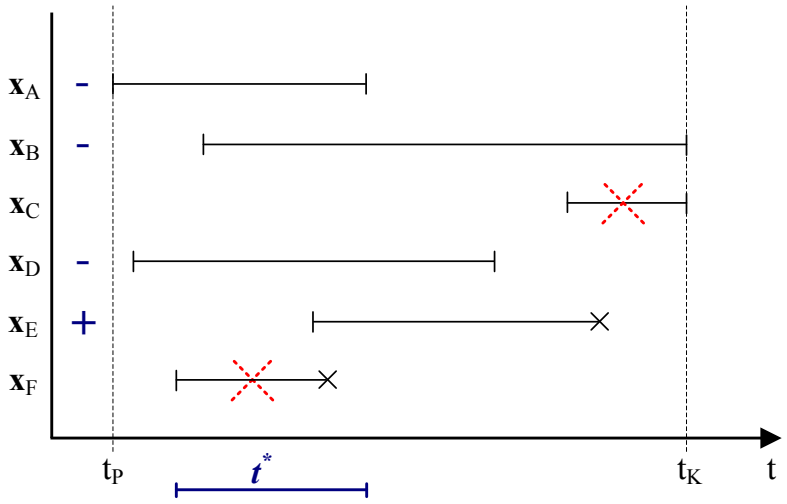
življenja cjelovitog uzorka), što ozbiljnije utječe na pristranost modela (Kattan *et al.*, 1998). Unatoč naivnoj pretpostavci, postupak je iznimno učinkovit za lakše cenzurirane podatke. Utjecaj cenzuriranih primjeraka pri njima se obično neutralizira običnim mehanizmima za suzbijanje pretreniranosti modela standardnih postupaka strojnog učenja.

U Snow *et al.* (1994) su se za učenje umjetne neuronske mreže za predviđanje ponavljanja karcinoma prostate nakon radikalne prostatektomije, cenzurirani pacijenti tretirali kao negativni, potpuno zanemarujući vrijeme praćenja. Takav model očito preferira negativne prognoze, stoga je možda pristran. U Štajduhar *et al.* (2009) napravljena je analiza utjecaja cenzure na učenje Bayesovih mreža tretiranjem cenzuriranih primjeraka kao negativnih. Opsežnom simulacijskom studijom pokazali su da algoritmi za učenje Bayesovih mreža učinkovito rukuju lakše cenzuriranim podacima, dok na višim razinama cenzure, ti algoritmi postaju neupotrebljivi.

3.2 Odstranjivanje prekratko praćenih primjeraka

Sljedeći pristup prilagodbe podataka odstranjuje sve primjerke koji su bili praćeni nedovoljno dugo. Odstranjivanjem prekratko praćenih primjeraka očekuje se odstranjivanje dviju vrsta primjeraka: 1) onih nereprezentativnih za domenu (npr. smrt izazvana prometnom nesrećom pacijenta koji sudjeluje u kliničkoj studiji povrata karcinoma prostate) te 2) onih reprezentativnih za domenu, ali s nepoznatim (cenzuriranim) ishodom (u slučaju da je količina podataka dovoljno velika). Odabir vrijednosti najmanjeg potrebnog vremena praćenja t^* ovisi o samim podacima. Postupak je predstavljen na slici 3.2.

Ukoliko je cenzuriranje u podacima slučajno, ovaj postupak gubi svaki smisao; tada se, naime, odstranjivanjem dijela podataka ne poboljšava kvaliteta preostalih podataka.



Slika 3.2: Primjer odstranjivanja prekratko praćenih primjeraka. Svi primjerci praćeni kraće od t^* odstranjuju se iz podataka ($\mathbf{x}_\mathrm{C}$ i $\mathbf{x}_\mathrm{F}$). Preostali cenzurirani primjerci tretiraju se kao negativni ($\mathbf{x}_\mathrm{A}$, $\mathbf{x}_\mathrm{B}$ i $\mathbf{x}_\mathrm{D}$).

Ovaj postupak pokazuje naklonost građenju pristranih modela, kada je udio cenzure u podacima visok (Kattan *et al.*, 1998; Ripley & Ripley, 2001).

U Delen *et al.* (2005) napravljena je usporedba performansi triju postupaka strojnog učenja na velikoj domeni proizašloj iz studije povrata raka dojke. Za modeliranje su koristili umjetne neuronske mreže, stabla odlučivanja i logističku regresiju. Iz skupa podataka isključili su sve zapise o pacijentima koji su bili praćeni kraće od pet godina. U jednom od testova na visokocenzuriranoj domeni o malignom melanomu, Ripley & Ripley (2001) isključili su sve zapise o pacijentima praćenim kraće od 2500 dana. Sličan je postupak korišten u Burke *et al.* (1997) u kojem su iz podataka dobivenih kliničkim studijama o kolorektalnom karcinomu i karcinomu dojke, iz učenja isključeni samo negativni pacijenti s prekratkim praćenjem (cenzurirani). Takvim je pristupom stvoren, očito, pristran model, koji preferira pozitivne prognoze.

3.3 Podjela na vremenske intervale

Negativan utjecaj cenzure povećava se s povećanjem vremena praćenja. Kada bi vrijeme praćenja bilo kraće, utjecaj cenzure bio bi smanjen te bi, posljedično, iz takvih podataka bilo lakše učiti. Vremenska se dimenzija praćenja T podijeli na k disjunktnih intervala, određenih vektorom granica $(t_1, \ldots, t_k, \infty)$, u kojem je $t_1 = 0$. Za svaki se vremenski interval izabere podskup podataka $\mathcal{D}_i \subseteq \mathcal{D}$ sastavljen iz primjeraka $\mathbf{x}$, za koje vrijedi $T(\mathbf{x}) \geq t_i$. Pritom se svaki, pojedinom podskupu D_i dodani primjerak, transformira na sljedeći način: ukoliko za dani $\mathbf{x}$ do trenutka t_{i+1} nije bio zabilježen ishod, ili je pak u nekom trenutku između t_i i t_{i+1} bio cenzuriran, tretira se kao negativan; u protivnom se

tretira kao pozitivan jer je u tom intervalu bio zabilježen ishod. U svakom od podskupova $\mathcal{D}_i$, cenzurirani primjerci i oni kojima u tom vremenskom intervalu nije prekinuto praćenje, tretiraju se kao negativni; primjerci kod kojih je pak bio zabilježen ishod u tom se intervalu tretiraju kao pozitivni. Ukupan broj primjeraka pridruženih svakom podskupu $\mathcal{D}_i$ manji je od broja primjeraka pridruženih podskupu $\mathcal{D}_{i-1}$. Postupak je ilustriran dvama primjerima na slici 3.3.

Podjelom skupa cenzuriranih podataka o preživljenju na k podskupova, grupiranih po bliskom vremenu praćenja, problem učenja modela pretvara se u k neovisnih problema učenja modela. Svaki od naučenih modela $\mathcal{M}_i$ moći će klasificirati neki novi primjerak $\mathbf{x}$, to jest bit će sposoban predviđati vjerojatnost preživljenja do vremena t_i. Konačna aposteriorna distribucija vjerojatnosti ishoda nepoznatog primjerka $\mathbf{x}$ računa se sljedećim izrazom:

$$\mathrm{P}(O|\mathbf{x},\mathcal{M}) = \frac{\sum_{i=1}^{k}\mathrm{P}(O|\mathbf{x},\mathcal{M}_i)}{k}, \tag{3.1}$$

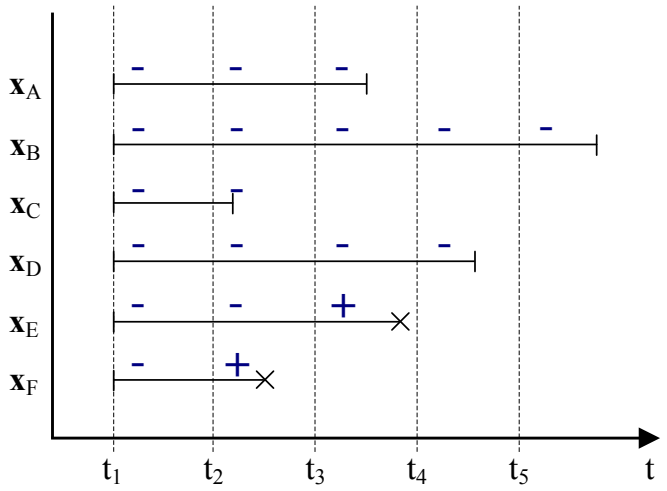
to jest kao aritmetička sredina aposteriornih distribucija vjerojatnosti svih modela. Opisani postupak odlučivanja posjeduje elemente tzv. učenja zajednicom modela (engl. *ensemble learning*) (Polikar, 2006).

U Jerez-Aragonés *et al.* (2003) i Lisboa *et al.* (2003), za učenje iz različitih domena kliničkih studija koje su istraživale povrat karcinoma dojke, bili su korišteni svi primjerci iz podataka (osim onih s nedostajućim vrijednostima). Primjerci su zatim bili razdijeljeni na zasebne skupove podataka za učenje, po jedan za svaki vremenski interval, iz kojih su se potom naučile umjetne neuronske mreže. Drugačiji je pristup opisan u Biganzoli *et al.* (1998), u kojem se iz svih podataka gradi jedna neuronska mreža; ona pak koristi informaciju o pripadnosti primjerka nekom vremenskom intervalu kao dodatni ulaz za modeliranje ispada.

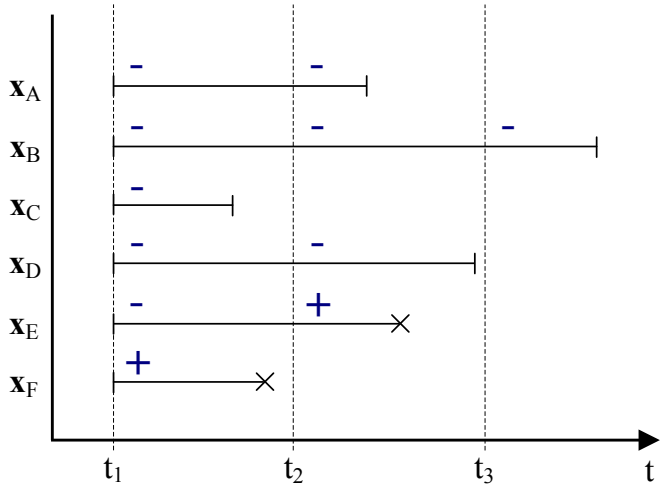
3.4 Podvajanje cenzuriranih primjeraka uz težinske faktore

Ovaj postupak proizlazi iz sljedeće tvrdnje: primjerak koji je kraće praćen ima manju apriornu vjerojatnost preživljenja od duže praćenog primjerka. Za svaki bi se poznati cenzurirani primjerak $\mathbf{x}^c$, stoga dalo odrediti, da je njegov ishod negativan s nekom vjerojatnošću $\mathrm{P}(O^-|\mathbf{x}^c)$ i pozitivan s vjerojatnošću $\mathrm{P}(O^+|\mathbf{x}^c) = 1 - \mathrm{P}(O^-|\mathbf{x}^c)$. Ta se informacija može iskoristiti podvajanjem svakog cenzuriranog primjerka u dva nova primjerka: jedan negativan s težinskim faktorom $w(\mathbf{x}^c) = \mathrm{P}(O^-)$ te jedan pozitivan s težinskim faktorom $w(\mathbf{x}^c) = 1 - \mathrm{P}(O^-)$.

Postupak predložen u Zupan *et al.* (2000) dodatno čini razliku između primjeraka cenzuriranih prije i nakon nekog trenutka t^*. Pretpostavlja se da su cenzurirani primjerci praćeni duže od t^* najvjerojatnije negativni. Uporište za to nalazi se u karakteristikama realnih domena, npr. u kliničkim studijama koje se bave proučavanjem povrata neke bolesti, u kojima se nakon nekoliko mjeseci ili godina praćenja, s velikom sigurnošću može reći je li primijenjeni tretman bio uspješan ili nije. U onkološkim studijama

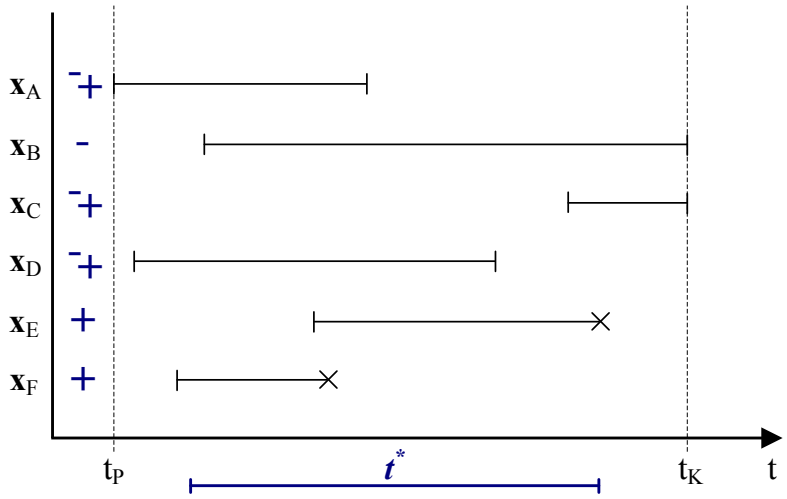


(a) Podjela na 5 vremenskih intervala



(b) Podjela na 3 vremenska intervala

Slika 3.3: Primjer podjele uzorka na više vremenskih intervala, sukladno vremenu praćenja. Svaki podskup uzorka koji odgovara nekom vremenskom intervalu posebno se koristi pri učenju. Primjerci cenzurirani u nekom intervalu u njemu se tretiraju kao negativni, kao i u svim prethodnim intervalima.



Slika 3.4: Primjer podvajanja svakog cenzuriranog primjerka praćenog kraće od t^* ($\mathbf{x}_\mathrm{A}$, $\mathbf{x}_\mathrm{C}$ i $\mathbf{x}_\mathrm{D}$) na pozitivnog i negativnog, s odgovarajućim težinskim faktorima. Primjerak $\mathbf{x}_\mathrm{B}$ je praćen duže od t^*, stoga se tretira kao negativan.

vezanim uz karcinom dojke, npr. najčešće se govori o periodu izlječenja $t^* = 5$ godina (Delen *et al.*, 2005). Postupak je ilustriran na slici 3.4.

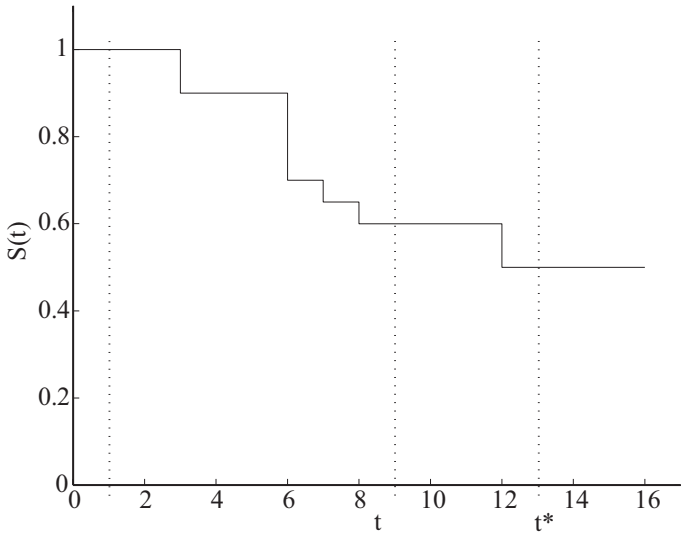
Težinski se faktori uzimaju u obzir prilikom učenja iz takvih podataka, što je podržano od strane mnoštva alata za strojno učenje. Procjena apriorne vjerojatnosti preživljenja primjerka cenzuriranog u trenutku t računa se omjerom vrijednosti procjene funkcije preživljenja Kaplan - Meier $\hat{S}_0(t)$ (Kaplan & Meier, 1958) u trenutku t i u trenutku t^* (pretpostavlja se da su svi primjerci praćeni dulje od t^* negativni). Procjena je određena izrazom:

$$\mathrm{P}(O^-|t) = \frac{\hat{S}_0(t)}{\hat{S}_0(t^*)}. \tag{3.2}$$

S obzirom na nepostojanje jasnog kriterija izbora vrijednosti t^* bez prethodnog ekspertnog znanja, a uzevši u obzir oblike procjena funkcija preživljenja Kaplan - Meier u simulacijskim studijama i na većini obrađenih realnih domena, u ovom je radu za t^* odabrano najduže opaženo vrijeme praćenja. Drugim riječima, $\hat{S}_0(t^*)$ najmanja je procijenjena vjerojatnost preživljenja što bi značilo da sigurnih negativnih primjeraka nema.

Izračun procjene apriorne vjerojatnosti preživljenja ilustriran je primjerom na slici 3.5. Kao vrijeme izlječenja odabrano je $t^* = 13$. Vjerojatnost preživljenja do vremena izlječenja u trenutku $t = 1$ jednaka je $\mathrm{P}(O^-|t = 1) = S(t^*)/S(t) = S(13)/S(1) = 0.5/1 = 0.5$; za trenutak $t = 9$ ona iznosi $\mathrm{P}(O^-|t = 9) = S(t^*)/S(t) = S(13)/S(9) = 0.5/0.6 = 0.8\overline{3}$ i vrijedi $\mathrm{P}(O^-|t = 9) > \mathrm{P}(O^-|t = 1)$.

U Zupan *et al.* (2000) opisana je analiza slučaja (engl. *case study*) povrata karci-



Slika 3.5: Primjer određivanja apriorne vjerojatnosti preživljenja omjerom vrijednosti u funkciji preživljenja $S(t)$. Vrijednost t^* predstavlja pretpostavljeni trenutak izlječenja.

noma prostate. Primijenili su opisani postupak na podatke o predoperativnim i podatke o postoperativnim pacijentima oboljelim od karcinoma prostate, koristeći vrijednost praga izlječenja $t^* = 7$ godina. Modeliranje stablima odlučivanja i naivnim Bayesovim klasifikatorom usporedili su s metodom proporcionalnih hazarda. U Ripley & Ripley (2001) koristi se sličan pristup - u ovom slučaju za učenje umjetnih neuronskih mreža iz domene o uznapredovalom karcinomu dojke i visokocenzurirane domene o malignom melanomu, ali isključivo na podacima za testiranje.

3.5 Odstranjivanje šuma cenzure

Ovaj je pristup prilagodbe podataka različit od ostalih predstavljenih time što koristi mehanizme strojnog učenja za ispravljanje početne klasifikacije nekih cenzuriranih primjeraka. Postupak temelji na činjenici da su pozitivni i negativni primjerci uzorkovani iz različitih inherentnih distribucija. Kada ne bi bili, postupak klasifikacije bio bi nemoguć. Odstranjivanje šuma cenzure ima zadatak pronaći sve cenzurirane primjerke uzorkovane iz inherentne distribucije pozitivnih primjeraka i promijeniti im klasifikaciju iz cenzurirane u pozitivnu. Opisani postupak predstavlja doprinos ove disertacije.

U nastavku je opisan postupak odstranjivanja šuma cenzure na osnovi izglednosti podataka (engl. *data likelihood*) u Bayesovim mrežama, kao jedan od mogućih načina pripremne obrade podataka odstranjivanjem šuma cenzure. Postupak je potekao iz Štajduhar & Bratko (2007), gdje se koristio za učenje standardnih problema, izvan

analize preživljenja. Početni skup podataka $\mathcal{D}$ podijeli se na dva disjunktna podskupa, $\mathcal{D}^+$ i $\mathcal{D}^c$, od kojih prvi sadrži sve pozitivne primjerke, a drugi sve cenzurirane. Iz oba se podskupa nauče zasebne Bayesove mreže, $\mathcal{B}^+$ i $\mathcal{B}^c$, korištenjem jednog od algoritama za učenje Bayesovih mreža iz podataka (za detalje pogledaj poglavlje 4). S obzirom na to da informacija o ishodu u oba podskupa podataka više ne služi svrsi, slobodno se prije učenja može ukloniti iz podataka, $\mathcal{D}^+ \leftarrow \mathcal{D}^+ \backslash \{O\}$, $\mathcal{D}^c \leftarrow \mathcal{D}^c \backslash \{O\}$.

Oba naučena modela trebala bi u nekoj mjeri biti dobra u opisivanju sebi inherentne distribucije - u većoj mjeri model, koji opisuje pozitivne primjerke (B^+), u manjoj mjeri model, koji opisuje cenzurirane primjerke (B^c). Sposobnost dobrog opisivanja modela B^c neposredno je ovisna o razini cenzure u podacima - što je ta manja, model će biti bliže inherentnoj distribuciji negativnih primjeraka (P^-); što je veća, veći će biti utjecaj mješavine inherentnih distribucija pozitivnih (P^+) i negativnih (P^-) primjeraka. Mjera kojom određujemo vjerojatnost da je određeni primjerak $\mathbf{x}$ bio uzorkovan iz neke distribucije P, je izglednost primjerka $\mathcal{L}(\mathbf{x}|\mathrm{P})$. Procjene izglednosti $\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^+)$ i $\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^-)$ određene su izrazima:

$$\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^+) = \mathcal{L}(\mathbf{x}|\mathcal{B}^+) = \mathrm{P}(\mathbf{x}|\mathcal{G}^+, \mathrm{P}^+) = \prod_{i=1}^{n} \mathrm{P}^+(V_i|\boldsymbol{\pi}(V_i)), \tag{3.3}$$

$$\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^-) \simeq \mathcal{L}(\mathbf{x}|\mathcal{B}^c) = \mathrm{P}(\mathbf{x}|\mathcal{G}^c, \mathrm{P}^c) = \prod_{i=1}^{n} \mathrm{P}^c(V_i|\boldsymbol{\pi}(V_i)), \tag{3.4}$$

u kojima je $\mathcal{B}^+ = (\mathcal{G}^+, \mathrm{P}^+)$ Bayesova mreža naučena iz pozitivnih primjeraka koju tvore graf $\mathcal{G}^+$ i združena distribucija vjerojatnosti P^+, dok je $\mathcal{B}^c = (\mathcal{G}^c, \mathrm{P}^c)$ Bayesova mreža naučena iz cenzuriranih primjeraka koju tvore graf $\mathcal{G}^c$ i združena distribucija vjerojatnosti P^c. Obje su mreže sastavljene iz skupa čvorišta $\mathbf{V} = \{V_1, \ldots, V_n\}$ (za detalje pogledaj poglavlje 4). Ukoliko razina cenzuriranja nije pretjerano visoka (ako je udio stvarnih pozitivnih primjeraka u nazivno cenzuriranim relativno malen), za cenzurirani će primjerak koji gravitira distribuciji P^- vrijediti $\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^-) > \hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^+)$, i obrnuto, za onog koji gravitira distribuciji P^+ vrijedit će $\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^+) > \hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^-)$. Što je primjerak sličniji inherentnoj distribuciji, bila ona pozitivna ili negativna, omjer njegovih procjena izglednosti na naučenim modelima bit će veći. Što je primjerak neutralniji, odnosno što je distribucija P^c više miješana podjednakom zastupljenošću P^+ i P^- (viša razina cenzure), to će isti omjer biti bliže vrijednosti 1. Ukoliko je razina cenzure u podacima veoma visoka, vrijedi $\mathrm{P}^c \simeq \mathrm{P}^+$, čime opisani postupak gubi svaki smisao.

Usporedbom izračunatih procjena izglednosti primjerka, može se odrediti spada li neki cenzurirani primjerak u podskup pozitivnih ili negativnih podataka. Korištenjem Bayesovog teorema, određenog izrazom:

$$\mathrm{P}(O|\mathbf{X}) = \mathrm{P}(O) \cdot \frac{\mathrm{P}(\mathbf{X}|O)}{\mathrm{P}(\mathbf{X})}, \tag{3.5}$$

moguće je odrediti aposteriornu distribuciju vjerojatnosti ishoda nekog cenzuriranog primjerka $\mathbf{x}$, na osnovi njemu procijenjenih izglednosti na oba modela. Ta se računa sljedećim izrazima:

$$\mathrm{P}(O^+|\mathbf{x}) = \hat{\mathrm{P}}(O^+) \cdot \frac{\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^+)}{\mathrm{P}(\mathbf{x})}, \tag{3.6}$$

$$\mathrm{P}(O^-|\mathbf{x}) = \hat{p}(O^-) \cdot \frac{\hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}^-)}{\mathrm{P}(\mathbf{x})}. \tag{3.7}$$

S obzirom na to da vrijedi $\mathrm{P}(O|\mathbf{x}) = \mathrm{P}(O^+|\mathbf{x}) + \mathrm{P}(O^-|\mathbf{x}) = 1$, pri izračunu aposteriorne distribucije vjerojatnosti, normalizacijom se može izostaviti član $\mathrm{P}(\mathbf{x})$ i dobiva se izraz:

$$\mathrm{P}(O|\mathbf{x}) = \hat{\mathrm{P}}(O) \cdot \hat{\mathcal{L}}(\mathbf{x}|\mathrm{P}) = \hat{\mathrm{P}}(O) \cdot \prod_{i=1}^{n} \mathrm{P}(V_i|\boldsymbol{\pi}(V_i)), \tag{3.8}$$

u kojem je $\mathrm{P} = \{\mathrm{P}^+, \mathrm{P}^-\}$. Apriorna distribucija vjerojatnosti ishoda računa se postupkom podvajanja cenzuriranih primjeraka uz težinske faktore (izraz 3.2). Postupak je detaljno opisan u sekciji 3.4.

Opreznim izborom praga ispravka cenzurirane klasifikacije P_c moguće je precizno odrediti koliko primjeraka želimo preklasificirati iz cenzuriranih u pozitivne. Ukoliko je $\mathrm{P}(O^+|\mathbf{x}) \geq \mathrm{P}_c$, cenzurirani primjerak postaje pozitivan, dok u protivnom postaje negativan. Nakon takve pripremne obrade podataka $\mathcal{D} = \{\mathcal{D}^+, \mathcal{D}^c\}$, dobiva se izmijenjeni skup podataka $\tilde{\mathcal{D}} = \{\tilde{\mathcal{D}}^+, \tilde{\mathcal{D}}^-\}$, koji se zatim bez ikakvih dodatnih promjena i bez informacije o vremenu praćenja, koristi za daljnje učenje modela standardnim metodama strojnog učenja.

U ovom je radu u svim simulacijskim studijama i na realnim domenama bio korišten prag $\mathrm{P}_c = 0.5$. Za učenje struktura $\mathcal{B}^+$ i $\mathcal{B}^c$ korištena je pohlepna metoda penjanja uzbrdo, opisana u sekciji 4.3, dok je učenje parametara mreža zasnovano na postupku najveće izglednosti, opisanom u sekciji 4.2.

Poglavlje 4

Učenje Bayesovih mreža

U ovom su poglavlju predstavljeni modeli predstavljanja znanja i mehanizmi njihovog učenja, korišteni u ovoj disertaciji. U sekciji 4.1 detaljnije su predstavljeni neovisnost i uzročnost, kao temeljni koncepti predstavljanja i interpretacije znanja Bayesovim mrežama, koji se logično nadovezuju na opis postupka zaključivanja, predstavljen u sekciji 2.2. Opis mehanizma učenja parametara, to jest lokalnih distribucija uvjetnih vjerojatnosti, uz poznatu strukturu Bayesove mreže, opisan je u sekciji 4.2, dok su algoritmi za učenje strukture Bayesove mreže opisani u sekciji 4.3. Slijedi opis ostalih modela predstavljanja znanja (sekcija 4.4), koji su bili korišteni u ovoj disertaciji radi dodatne provjere učinkovitosti korištenih postupaka prilagodbe cenzuriranih podataka o preživljenju.

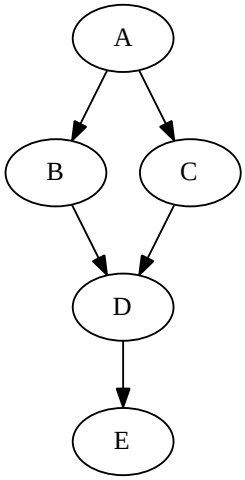
4.1 Notacija i temeljni koncepti

U Pearl (1988) je Bayesova mreža $\mathcal{B}$ formalno određena parom $\mathcal{B} = (\mathcal{G}, \mathrm{P}(\mathcal{G}))$, gdje je $\mathcal{G}$ usmjeren aciklički graf $\mathcal{G} = (\mathbf{V}(\mathcal{G}), \mathbf{A}(\mathcal{G}))$ sa skupom čvorišta $\mathbf{V}(\mathcal{G}) = \{V_1, V_2, \ldots, V_n\}$, koji predstavljaju slučajne kovarijate $\mathbf{X}$ i varijatu od interesa O, te skupom lukova $\mathbf{A}(\mathcal{G}) \subseteq \mathbf{V}(\mathcal{G}) \times \mathbf{V}(\mathcal{G})$, koji predstavljaju uvjetne ovisnosti među tim čvorištima. Nad skupom čvorišta $\mathbf{V}$ određena je združena distribucija vjerojatnosti $\mathrm{P}(\mathcal{G})$, koja uvažava (ne)ovisnosti predstavljene grafom: $\mathrm{P}(V_1, \ldots, V_n) = \prod_{i=1}^{n} \mathrm{P}(V_i|\boldsymbol{\pi}(V_i))$, gdje $\boldsymbol{\pi}(V_i)$ označava kovarijate koje odgovaraju roditeljima čvorišta V_i.

4.1.1 Interpretacija neovisnosti pravilima d-razdvajanja

Pravila d-razdvajanja (engl. *d-separation*) koristimo za određivanje neovisnosti u nekoj domeni iz strukture Bayesove mreže. Pretpostavimo da se čvorište V_Z nalazi negdje na neusmjerenom putu $\mathcal{P}$ između čvorišta V_X i V_Y te da je $\mathcal{P}$ jedini put između V_X i V_Y u grafu $\mathcal{G}$. Htjeli bismo odrediti je li V_X neovisan od V_Y, ako nam je poznata vrijednost V_Z.

U Pearl (1988) je postupak d-razdvajanja opisan na sljedeći način. Neka su $\mathbf{S}$, $\mathbf{T}$ i $\mathbf{V}$ tri disjunktna podskupa čvorišta u usmjerenom acikličkom grafu $\mathcal{G}$. Neka je $\mathcal{P}$ bilo



Slika 4.1: Primjer grafa i d-razdvajanja.

koji put među čvorištima u **S** i čvorištima u **T**, neovisno od smjera pojedinih lukova među susjednim čvorištima. Možemo reći da **V** blokira $\mathcal{P}$, ako na putu postoji čvorište V_Z, za koje vrijedi jedan od uvjeta:

- V_Z ima konvergirajuće lukove na $\mathcal{P}$ i vrijedi, da niti V_Z niti jedan od njegovih potomaka nije u **V**, ili
- V_Z nema konvergirajuće lukove na $\mathcal{P}$ i nalazi se u **V**.

V d-razdvaja **S** od **T**, što pišemo ($\mathbf{S} \perp_{\mathcal{G}} \mathbf{T} \mid \mathbf{V}$), ako i samo ako **V** blokira sve puteve od čvorišta u **S** do čvorišta u **T**. Čvorište V_Z ima konvergirajuće lukove (engl. *converging arcs*) na $\mathcal{P}$ ako su oba njemu neposredna luka na $\mathcal{P}$ usmjerena k njemu $V_X \rightarrow V_Z \leftarrow V_Y$.

Postupak određivanja neovisnosti na osnovi d-razdvajanja ilustriran je na primjeru grafa sa slike 4.1 (primjer preuzet iz Krause, 1998). Neka je $\mathbf{S} = \{V_B\}$, $\mathbf{T} = \{V_C\}$ i, za početak, $\mathbf{V} = \{V_A\}$. Potrebno je odrediti vrijedi li $\mathbf{S} \perp_{\mathcal{G}} \mathbf{T} \mid \mathbf{V}$. Prvo je potrebno odrediti sve moguće puteve $\mathcal{P}$ između **S** i **T**. Postoje dva puta između V_B i V_C, konkretno $V_B \leftarrow V_A \rightarrow V_C$ i $V_B \rightarrow V_D \leftarrow V_C$. U prvom primjeru ni jedno čvorište u **V** nema konvergirajuće lukove. U drugom primjeru čvorište s konvergirajućim lukovima i njegov potomak nisu u **V**. Očito je da $\mathbf{V} = \{V_A\}$ zadovoljava oba uvjeta i zato možemo tvrditi da V_A d-razdvaja V_B od V_C na grafu $\mathcal{G}$, odnosno $\{V_B\} \perp_{\mathcal{G}} \{V_C\} \mid \{V_A\}$. Što se dogodi ako je $\mathbf{V} = \{V_A, V_E\}$? Čvorište V_E (potomak čvorišta V_D) sada je uključeno u **V**. Ako poznajemo vrijednost čvorišta V_E, njegovi uzroci (čvorišta V_B i V_C) postat će međusobno neovisni. Skup $\{V_A, V_E\}$ ne ispunjava uvjete te vrijedi da V_B i V_C nisu d-razdvojeni skupom $\{V_A, V_E\}$, odnosno $\neg(\{V_B\} \perp_{\mathcal{G}} \{V_C\} \mid \{V_A, V_E\})$.

Usmjeren aciklički graf je preslika neovisnosti (engl. *independency map*, I-map) distribucije vjerojatnosti P, ako sve uvjetne neovisnosti sadržane u $\mathcal{G}$ vrijede u P po pravilima d-razdvajanja. Jednostavnije rečeno, za svaki **S**, **T** i **V** vrijedi ($\mathbf{S} \perp_{\mathcal{G}} \mathbf{T} \mid \mathbf{V}$) $\Rightarrow$

($\mathbf{S} \perp_{\mathrm{P}} \mathbf{T} \mid \mathbf{V}$), gdje ($\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$) predstavlja d-razdvajanje $\mathbf{S}$ i $\mathbf{T}$ glede $\mathbf{V}$, dok ($\mathbf{S} \perp_{\mathrm{P}} \mathbf{T} \mid \mathbf{V}$) označava uvjetnu neovisnost u distribuciji P. $\mathcal{G}$ je Bayesova mreža ako, i samo ako je minimalna preslika neovisnosti iz P, odnosno ako nije moguće odstraniti bilo koji luk iz $\mathcal{G}$ bez da bismo time negirali karakteristike preslike neovisnosti (Borgelt & Kruse, 2002; Pearl, 1988).

Iz strukture grafa je, dakle, vidljiv skup relacija uvjetnih neovisnosti na uključenim čvorištima. Da bi model bio ispravan, potrebne su sve relacije. Po drugoj strani vrijedi, da d-povezana (odnosno ne d-razdvojena) čvorišta u grafu $\mathcal{G}$ nisu nužno ovisna o distribuciji P. Npr. potpuno povezan usmjeren aciklički graf, uvijek je preslika neovisnosti (to znači da nema neovisnosti), no ako slijedimo definiciju, ne predstavlja ispravnu Bayesovu mrežu jer nije minimalan.

4.1.2 Uzročnost

Pored opisa neovisnosti, strukturu grafa Bayesove mreže u nekim domenama možemo koristiti i za prikaz uzročno-posljedičnih odnosa, kroz lukove i njihove smjerove. U takvim primjerima roditelj čvorišta predstavlja neposredni uzrok sadržaja koji predstavlja to čvorište. Ova izjava je istinita samo u slučaju da vrijede sljedeće pretpostavke (Borgelt & Kruse, 2002; Pearl, 1988):

- Ako ne postoje zajednički neopaženi uzroci dvaju ili više opažanih čvorišta u domeni, onda vrijedi pretpostavka uzročne dovoljnosti (engl. *causal sufficiency*). Neopažana čvorišta zovemo latentna ili skrivena čvorišta.
- Glede uzročne dovoljnosti postoji mogućnost da više od jedne strukture mreže odgovara ograničenjima koja proizlaze iz domene. Ta ograničenja su statističke neovisnosti izvedene iz podataka. Samo jedna od tih mreža može biti istinita na uzročnom modelu koji predstavlja prave uzročno-posljedične relacije, koje upravljaju mehanizmom nastajanja podataka u domeni.

Bayesove mreže možemo interpretirati kao uzročne modele s običnim uzročno-posljedičnim tumačenjem samo u primjeru kada znamo da gornje pretpostavke vrijede (što je rijetko, pogotovo za prvu pretpostavku).

4.2 Učenje lokalnih distribucija uvjetnih vjerojatnosti

Drugi dio Bayesove mreže je, pored grafa $\mathcal{G}$, skup lokalnih distribucija uvjetnih vjerojatnosti (engl. *conditional probability table*). Na strukturi grafa $\mathcal{G}$ on predstavlja združenu distribuciju vjerojatnosti u domeni (engl. *joint probability distribution*). Združena distribucija vjerojatnosti P određena je izrazom:

$$\mathrm{P}(V_1, \ldots, V_n) = \prod_{i=1}^{n} \mathrm{P}(V_i | \boldsymbol{\pi}(V_i)), \tag{4.1}$$

u kojem je $\boldsymbol{\pi}(V_i)$ skup roditelja čvorišta V_i. Iz izraza 4.1 očito slijedi da se P može faktorizirati na manje, lokalne distribucije uvjetnih vjerojatnosti, po jednu za svako

čvorište i njegove roditelje. Lokalne distribucije uvjetnih vjerojatnosti dovoljne su za rekonstrukciju združene distribucije vjerojatnosti u domeni, ako ih pomnožimo po pravilima uvjetnih neovisnosti (engl. *conditional independence rules*) (Pearl, 1988), koje su pak određene strukturom mreže $\mathcal{G}$.

Lokalnu distribuciju uvjetnih vjerojatnosti može se predstaviti bilo kojom porodicom distribucija vjerojatnosti. Neovisnosti predstavljene strukturom Bayesove mreže vrijede za svakog člana porodice, koji je s njom spojiv, to jest vrijede za bilo kakav izbor parametara lokalnih distribucija uvjetnih vjerojatnosti. Ako su neko čvorište i njegov roditelj u grafu diskretni, lokalna je distribucija uvjetnih vjerojatnosti obično opisana multinomnom distribucijom (Bishop, 2007) te se predstavlja tablicama uvjetnih vjerojatnosti. Kad su čvorišta kontinuirana, koriste se linearni Gaussovi modeli.

Učenje parametara uz poznatu strukturu mreže dobro je definiran statistički problem. Na osnovi parametara (ako postoje i prije no što su upotrijebljeni podaci) pretpostavimo početnu distribuciju lokalnih tablica uvjetnih vjerojatnosti (u protivnom pretpostavimo da je distribucija uniformna). Poželjno je, da je početna distribucija pridruživa izvedenoj. Početna je distribucija pridruživa (engl. *conjugate*), kada aposteriorna distribucija na parametrima spada u istu porodicu kao i početna, ali s drugačijim hiperparametrima (parametri distribucije parametara). U ovom se radu za lokalne tablice uvjetnih vjerojatnosti koriste isključivo multinomne distribucije te su samo one predstavljene detaljnije.

Za multinomne distribucije dobijemo pridruživu početnu distribuciju iz Dirichletove porodice (Bishop, 2007). Takva distribucija je primjenjiva za bilo koju kardinalnost čvorišta (binarnim bi čvorištima odgovarala i beta distribucija). Vjerojatnost svake vrijednosti čvorišta V_i u lokalnoj distribuciji vjerojatnosti za kombinaciju roditelja $\boldsymbol{\pi}_{ij}$ označavamo s P_{ijk} za $k = 1, \ldots, r_i$, gdje je r_i broj vrijednosti diskretnog čvorišta V_i. Dirichletova distribucija određena je izrazom:

$$\mathrm{P}(\mathrm{P}_{ij1}, \mathrm{P}_{ij2}, \ldots, \mathrm{P}_{ijr_i}|\mathcal{G}) = Dir(\alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijr_i}) = \Gamma(\alpha_{ij}) \prod_{k=1}^{r_i} \frac{\mathrm{P}_{ijk}^{\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk})}, \qquad (4.2)$$

u kojem su α_{ijk} hiperparametri te vrijedi $\alpha_{ijk} \geq 0$, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. Vrijednost hiperparametara dodatno oblikuje apriornu distribuciju vjerojatnosti. Ako je apriorna distribucija vjerojatnosti poznata, tada vrijednost svakog hiperparametra α_{ijk} možemo odrediti brojnošću vrijednosti k čvorišta V_i, glede na kombinaciju vrijednosti njegovih roditelja $\boldsymbol{\pi}_{ij}$ (Bishop, 2007). Gama funkcija u formuli predstavlja normalizacijsku konstantu i definirana je izrazom $\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt$. Ako pretpostavimo neovisnost lokalnih i neovisnost globalnih parametara (obje pretpostavke neovisnosti proizlaze iz strukture mreže, pogledaj Heckerman, 1996; Cooper & Herskovits, 1992), distribucija na skupu parametara za cijelu Bayesovu mrežu određena je izrazom:

$$\mathrm{P}(\mathbf{P}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \Gamma(\alpha_{ij}) \prod_{k=1}^{r_i} \frac{\mathrm{P}_{ijk}^{\alpha_{ijk}-1}}{\Gamma(\alpha_{ijk})}, \qquad (4.3)$$

u kojem je q_i produkt kardinalnosti čvorišta u $\boldsymbol{\pi}_{ij}$. Aposteriorna je vjerojatnost na distribuciji također član Dirichletove porodice i vrijedi:

$$\mathrm{P}(\mathrm{P}_{ij1}, \mathrm{P}_{ij2}, \ldots, \mathrm{P}_{ijr_i}|\mathcal{G}, \mathcal{D}) = Dir(N_{ij1} + \alpha_{ij1}, N_{ij2} + \alpha_{ij2}, \ldots, N_{ijr_i} + \alpha_{ijr_i}), \quad (4.4)$$

$$\mathrm{P}(\mathbf{P}|\mathcal{G}, \mathcal{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \Gamma(N_{ij} + \alpha_{ij}) \prod_{k=1}^{r_i} \frac{\mathrm{P}_{ijk}^{N_{ijk}+\alpha_{ijk}-1}}{\Gamma(N_{ijk} + \alpha_{ijk})}, \quad (4.5)$$

gdje je N_{ijk} broj primjeraka iz $\mathcal{D}$ koji spada u skupinu k lokalne tablice uvjetnih vjerojatnosti za čvorište V_i, uz kombinaciju vrijednosti roditelja $\boldsymbol{\pi}_{ij}$. Za predviđanje bilo koje kombinacije $Q(V_1, \ldots, V_n)$ izračunamo srednje vrijednosti za sve moguće nepoznate parametre i dodamo im težinske faktore s izvedenom vjerojatnosti svake vrijednosti:

$$\mathrm{P}(Q(V_1, \ldots, V_n)|\mathcal{G}, \mathcal{D}) = \int Q(V_1, \ldots, V_n)\mathrm{P}(\mathbf{P}|\mathcal{G}, \mathcal{D})d\mathbf{P}. \quad (4.6)$$

Često se, radi jednostavnosti, umjesto cijele distribucije koriste samo parametri izračunati postupkom najveće izglednosti (engl. *maximum likelihood*, ML) (Borgelt & Kruse, 2002). Najveća procijenjena izglednost za P_{ijk} je:

$$\hat{\mathrm{P}}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}. \quad (4.7)$$

4.3 Učenje strukture mreže

Problem učenja Bayesove mreže može se postaviti na sljedeći način: za dani skup podataka $\mathcal{D}$, pronađi model $\mathcal{B}$ koji najbolje opisuje $\mathcal{D}$. Uobičajeni je pristup rješavanju ovog problema uvođenje kriterijske funkcije, koja će vrednovati svaki mogući model na temelju $\mathcal{D}$ te pronaći najbolju mrežu sukladno korištenoj metrici, kao npr. u Chickering (2002); Cooper & Herskovits (1992); Friedman & Koller (2003); Heckerman *et al.* (1995). Korištene kriterijske funkcije najčešće se temelje na funkciji vjerovanja (engl. *belief scoring functions*) (Heckerman *et al.*, 1995) te najmanjoj dužini opisa (engl. *Minimum Description Length*, MDL) (Lam & Bacchus, 1994). Alternativni pristup rješavanju problema učenja temelji se na ograničenjima (engl. *constraint-based learning*). Ta su ograničenja obično tvrdnje uvjetnih neovisnosti, određene statističkim pokusima nad podacima. Ovaj je pristup temeljito opisan u Cheng *et al.* (2002); Pearl (2000); Spirtes *et al.* (2000). Iz razloga što su pretraživanja prostora struktura grafova i prostora uvjetnih neovisnosti računski vrlo zahtjevni postupci, postoji mnogo njihovih adaptacija (Abellan *et al.*, 2006; Bromberg & Margaritis, 2009; Hruschka & Ebecken, 2007; Tsamardinos *et al.*, 2006; Xie & Geng, 2008).

Nakon što se odredi struktura Bayesove mreže, tablice uvjetnih vjerojatnosti se odrede neposredno iz podataka računanjem frekvencijskih distribucija na uvjetnim potprostorima (sekcija 4.2). Uz naučenu ili već postojeću Bayesovu mrežu, jednostavno je

predviđati vrijednosti varijate od interesa (ishoda) na osnovi dokaza. Za to se koriste posebni postupci probabilističkog zaključivanja, opisani u Pearl (1988).

Iako postoji pregršt postupaka za učenje strukture Bayesovih mreža iz podataka, velika većina njih spada u jednu od prije navedenih skupina - temelje se ili na ograničenjima, ili na uspjehu. Kako bi se pojednostavila usporedba između tih dvaju pristupa, iz svake je skupine izabran po jedan temeljni i dobro poznati algoritam. Iz skupine temeljene na uspjehu izabran je pohlepni algoritam penjanja uzbrdo (engl. *hill-climbing*), dok je iz druge skupine, one temeljene na ograničenjima, izabran algoritam uvjetnih neovisnosti (engl. *conditional independence*). Navedeni su algoritmi opisani u nastavku.

4.3.1 Algoritam uvjetnih neovisnosti

Algoritam uvjetnih neovisnosti (engl. *conditional independence*, CI) (Verma & Pearl, 1992) koristi pokuse uvjetnih neovisnosti kako bi pronašao strukturu Bayesove mreže, nakon čega obradom skupa određenih pravila određuje smjerove lukova.

Počevši s potpunim neusmjerenim grafom, algoritam pokušava pronaći uvjetne neovisnosti $\langle V_x, V_y | \mathbf{V}_Z \rangle$ u podacima. Za svaki par čvorišta $\{V_x, V_y\}$, algoritam razmatra skupove $\mathbf{V}_Z$ počevši s kardinalnošću nula, zatim jedan, sve do ukupnog broja čvorišta umanjenog za dva. Skup $\mathbf{V}_Z$ podskup je skupa čvorišta koja su susjedi čvorištima V_x i V_y. Ukoliko se utvrdi neovisnost, luk između V_x i V_y se uklanja iz strukture mreže. Provjera je li par čvorišta $\{V_x, V_y\}$ uvjetno neovisan za dani skup čvorišta $\mathbf{V}_Z$, izvodi se usporedbom strukture mreže s lukovima $\forall V_z \in \mathbf{V}_Z : V_z \rightarrow V_y$ s onom s lukovima $\{V_x \rightarrow V_y\} \cup \forall V_z \in \mathbf{V}_Z : V_z \rightarrow V_y$. Pokus se izvodi korištenjem Bayesove metrike (Heckerman *et al.*, 1995).

Po određivanju strukture mreže, potrebno je usmjeriti lukove. Za svaki nesusjedni par čvorišta $\{V_x, V_y\}$ u postavi $V_x - V_z - V_y$, ako vrijedi $V_z \in \mathbf{V}_Z$, tada usmjeri sve lukove na način $V_x \rightarrow V_z \leftarrow V_y$ (osim ako je jedan od lukova već usmjeren). Na koncu se za usmjeravanje preostalih neusmjerenih lukova primjeni skup grafičkih pravila, opisanih u Verma & Pearl (1992).

Algoritam pretpostavlja da skup podataka ima savršenu mapu (engl. *perfect map*). Graf $\mathcal{G}$ nazivamo savršenom mapom (Pearl, 1988) skupa ovisnosti Σ: 1) ako se svaka ovisnost, na koju logički ukazuje Σ, može zaključiti iz $\mathcal{G}$ te 2) ako je svaka ovisnost zaključena iz $\mathcal{G}$ logički proizašla iz Σ. Ukoliko navedena tvrdnja nije istinita, algoritam neće biti u stanju svakom otkrivenom luku dodijeliti smjer; iz tog se razloga mora oprezno koristiti.

U praksi se neusmjereni luk tretira kao da se sastoji od dva suprotno usmjerena luka. Iako takav tretman ne odgovara definiciji Bayesove mreže, može se koristiti iz sljedećih razloga. Prvi je taj, da neki luk, bio on usmjeren ili ne, predstavlja interakciju među čvorištima (iako u tom slučaju bez mogućeg objašnjenja uzročnosti). Drugi je pak taj, da se može zaključivati izračunom združene distribucije vjerojatnosti nad dokazima i naknadnom normalizacijom po procijenjenim distribucijama vjerojatnosti. Postoje i sofisticiraniji načini tretiranja neusmjerenih lukova, poput zabrane određivanja suprotno usmjerenih lukova na način da se posljedice neovisnosti obrađuju redom, ovisno o nji-

hovoj jakosti. Takvi postupci nisu bili korišteni u ovom radu, to jest neusmjereni lukovi su bili ostavljeni u svom izvornom obliku.

4.3.2 Pohlepna metoda penjanja uzbrdo

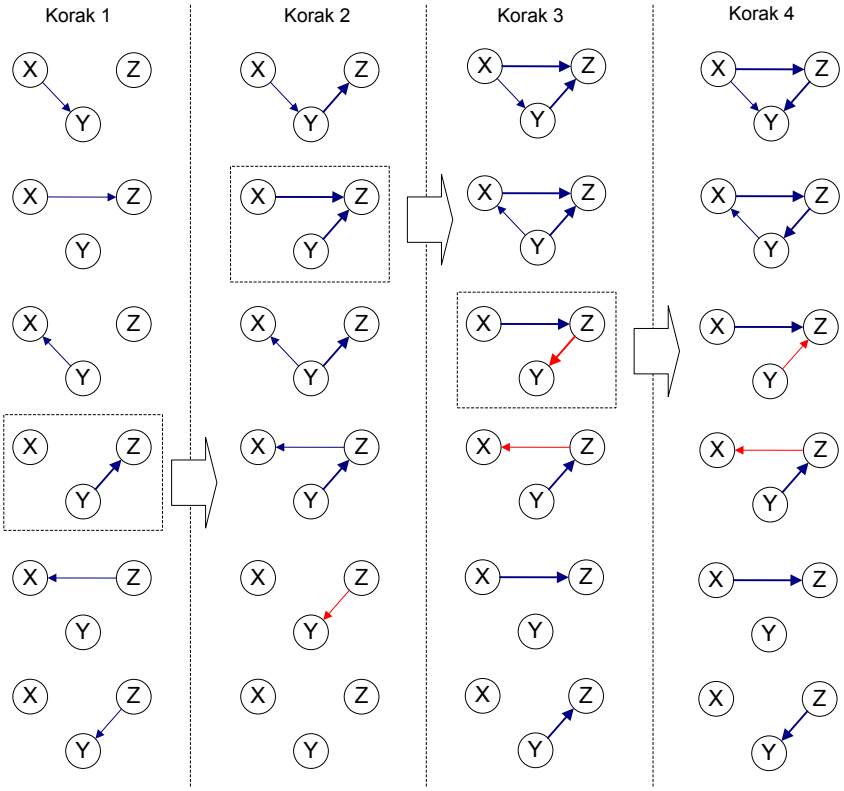
Metode temeljene na uspjehu dodjeljuju ocjenu uspješnosti svakoj Bayesovoj mreži kandidatu, obično neku koja mjeri koliko dobro ta Bayesova mreža opisuje dani skup podataka $\mathcal{D}$. Ocjena uspješnosti Bayesove mreže određene strukturom $\mathcal{G}$ i iz podataka $\mathcal{D}$ procijenjenim parametrima $\hat{\mathrm{P}}$ opisana je izglednošću podataka $\mathrm{P}(\mathcal{D}|\mathcal{G},\hat{\mathrm{P}})$. Kako bi se spriječila pretreniranost modela, ocjeni se dodaje faktor koji kažnjava odveć složene strukture (nalik ID3 podrezivanju stabla). Najmanja dužina opisa (MDL) (Grünwald *et al.*, 2005; Lam & Bacchus, 1994) koristi se kao kriterijska funkcija koju treba minimizirati. Određena je izrazom:

$$MDL(\mathcal{G},\mathcal{D}) = \frac{d}{2}\log_2 N - \log_2 \mathrm{P}(\mathcal{D}|\mathcal{G},\hat{\mathrm{P}}), \tag{4.8}$$

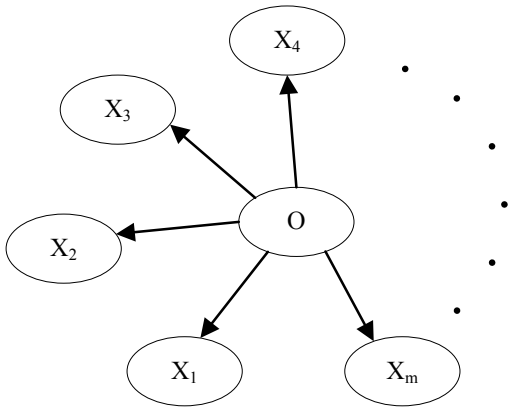
u kojem je d broj slobodnih parametara multinomnih lokalnih tablica uvjetnih vjerojatnosti, dok je N ukupan broj primjeraka u uzorku.

Pošto je prostor svih mogućih struktura u najmanju ruku eksponencijalan glede broja čvorišta n (postoji $n(n-1)/2$ mogućih neusmjerenih lukova te $2^{n(n-1)/2}$ mogućih struktura za svaki od podskupa tih lukova, pritom ne uzevši u obzir orijentacije lukova), pristup sirovom snagom, koji bi izračunao ocjenu uspješnosti svake strukture Bayesove mreže, nije primjenjiv ni na jednoj, osim na najjednostavnijoj domeni. Umjesto toga, obično se koriste heuristički algoritmi za pretraživanje, poput algoritma penjanja uzbrdo (engl. *hill-climbing*, HC) (Russell & Norvig, 2002).

Pretraživanje počinje praznim grafom. Za svaki par čvorišta algoritam provjeri učinak dodavanja, uklanjanja ili obrtanja luka na ocjenu uspješnosti modela. Postupak završava u trenutku kada više nema operacije nad jednim lukom koja bi mogla smanjiti vrijednost kriterijske funkcije. Na slici 4.2 ilustriran je rad algoritma na izmišljenom problemu. U svakom koraku algoritam na grafu iz prethodnog koraka odabere operaciju nad lukom, koja vrati model s najboljom ocjenom (npr. najnižim MDL-om). U prvom je koraku graf prazan (nema lukova) te je jedina moguća elementarna operacija dodavanje luka; najbolju ocjenu modelu donosi operacija $Y \to Z$. U drugom koraku se može dodati novi luk ili ukloniti, ili okrenuti, već postojeći ($Y \to Z$); izvođenjem elementarne operacije dodavanja $X \to Z$ dobijemo najbolji model $X \to Z \leftarrow Y$. U trećem su koraku ponovo moguće sve elementarne operacije; operacija koja nudi najveće poboljšanje postojećem modelu jest operacija okretanja postojećeg luka $Y \rightleftarrows Z$, što tvori novi model $X \to Z \to Y$. U četvrtom koraku postupak ne nalazi bolji model od onog nađenog u trećem koraku te se tu zaustavlja. Nažalost, ne postoji nikakvo jamstvo da će se algoritam zaustaviti u globalnom minimumu. Iako jednostavne perturbacije, poput višestrukog ponovnog pokretanja iz slučajnih početnih mreža ili simuliranog žarenja (engl. *simulated annealing*) (Janžura & Nielsen, 2006), mogu biti korištene kako



Slika 4.2: Primjer postupka otkrivanja strukture Bayesove mreže pohlepnim algoritmom penjanja uzbrdo.



Slika 4.3: Grafički prikaz naivnog Bayesovog klasifikatora.

bi se povećala vjerojatnost dostizanja globalnog minimuma, one u ovom radu nisu bile korištene.

4.4 Ostali modeli

U nastavku su predstavljena još dva modela koja su bila korištena za dodatnu provjeru učinkovitosti opisanih postupaka prilagodbe cenzuriranih podataka o preživljenju.

4.4.1 Naivni Bayesov klasifikator

Naivni Bayesov klasifikator (engl. *naive Bayes classifier*, NB) (Hand & Yu, 2001) jednostavan je probabilistički klasifikator, temeljen na snažnim pretpostavkama neovisnosti kovarijata. Parametri modela odrede se procjenom maksimalne izglednosti. Distribucija vjerojatnosti ishoda shodna dokazima $\mathbf{X}$ određena je izrazom:

$$P(O|\mathbf{X}) = P(O) \cdot \prod_{i=1}^{m} \frac{P(X_i|O)}{P(X_i)}, \tag{4.9}$$

u kojem je $P(O)$ apriorna vjerojatnost zbivanja događaja od interesa, dok je $P(X_i|O)$ uvjetna vjerojatnost kovarijate X_i, uz poznati ishod O. Slika 4.3 prikazuje naivni Bayesov klasifikator kao poseban oblik Bayesove mreže, s jedinim mogućim tipom luka - onim usmjerenim od varijate od interesa (uzrok) prema svakoj od kovarijata (posljedica). Unatoč naivnoj pretpostavci, naivni Bayesov klasifikator iznenađujuće je moćan klasifikator za modeliranje složenih realnih domena (Domingos & Pazzani, 1997). Karakteristike naivnog Bayesovog klasifikatora daju se jednostavno predočiti u obliku nomograma (Mozina *et al.*, 2004), što ovom modelu daje dodatnu prednost u potpori odlučivanju.

4.4.2 Model proporcionalnih hazarda

Vjerojatno najpopularnija metoda modeliranja preživljenja statistički je postupak zvan regresijska analiza proporcionalnih hazarda (PH), također poznata pod nazivom Coxova regresija (Cox, 1972). Coxova regresija modelira hazard u ovisnosti od vrijednosti kovarijata. Funkcija hazarda određena je izrazom:

$$h(t|\mathbf{X}) = h_0(t)e^{\boldsymbol{\beta}\mathbf{X}'}, \tag{4.10}$$

u kojem je $\mathbf{X}$ skup kovarijata, $\boldsymbol{\beta}$ vektor regresijskih koeficijenata te $h_0(t)$ funkcija osnovnog hazarda (engl. *baseline hazard*), koja se može interpretirati kao funkcija hazarda u slučaju u kojem su sve kovarijate jednake nuli. Ova metoda pretpostavlja da su učinci različitih kovarijata na funkciju preživljenja konstantni kroz vrijeme, što joj predstavlja manu u nekim slučajevima. Regresijski koeficijenti daju se odrediti korištenjem metode parcijalne izglednosti (engl. *partial likelihood*), što nam omogućuje zanemarivanje osnovnog hazarda. Coxov se model može tumačiti razmatranjem regresijskih koeficijenata pojedinačno: pozitivan koeficijent povećava hazard, dok ga negativan smanjuje, u iznosu veličine njegove apsolutne vrijednosti. Jednom kad je model naučen, funkcija preživljenja novog primjerka, temeljena na njegovim opaženim karakteristikama, može se odrediti izrazom:

$$S(t|\mathbf{X}) = S_0(t)^{exp(\boldsymbol{\beta}\mathbf{X}')}, \tag{4.11}$$

u kojem je $S_0(t)$ osnovna funkcija preživljenja (Lee & Wang, 2003):

$$S_0(t) = e^{-\int_0^t h_0(s)ds}. \tag{4.12}$$

Ukoliko se regresijski model proporcionalnih hazarda želi koristiti za klasifikaciju, funkciju preživljenja je nužno preslikati u distribuciju vjerojatnosti. To je na prvom mjestu neophodno radi usporedbe performansi Coxove regresije i postupaka strojnog učenja. U ovom je radu za predviđanje konačnog ishoda korištena vrijednost funkcije preživljenja u medijani vremena praćenja danog uzorka $P(O|\mathbf{X}) = S(t_{MED}|\mathbf{X})$.

Poglavlje 5

Opis eksperimentalnog vrednovanja postupaka prilagodbe podataka

5.1 Metrike vrednovanja

Mjerenjem razlike između predviđanog i stvarnog ishoda kod cenzuriranih podataka o preživljenju nije moguće ustanoviti stvarnu vrsnost nekog modela. Razlog tomu je da je informacija o ishodu, zabilježena u nekom skupu podataka, nepotpuna odnosno kontaminirana cenzuriranim primjercima gdje se ishod zaista zbio, no nije bio zabilježen zbog kratkog vremena praćenja. Čak i u slučaju da je udio cenzuriranih primjeraka u nekom skupu podataka relativno malen, pogreška u mjerenju može dovesti do pogrešnog zaključka. Logičan bi izbor bio odstraniti sve cenzurirane primjerke iz plana izvedbe testiranja, što nas dovodi do sljedećeg pitanja: gdje (ili zapravo, kada) se u vremenu praćenja nalazi granična vrijednost koja razdvaja primjerke stvarnih negativnih ishoda od onih s mogućim negativnim ishodom?

Ne postoji ni jedan najbolji, najtočniji ili najjednostavniji način vrednovanja postupaka za učenje modela za predviđanje preživljenja ili klasifikatora za predviđanje konačnog stvarnog ishoda. Iz tog je razloga u ovom radu korištena zajednica metrika vrednovanja, kako bi se interpretacijom različitih rezultata stvorila prava slika o vrsnosti predstavljenih postupaka. Procjena performansi postupaka modeliranja izvela se na korištenim podacima pomoću dviju skupina metrika vrednovanja: onih standardnih za strojno učenje te onih proizašlih iz analize preživljenja.

Tablica 5.1: Matrica konfuzije opisuje odnos ishoda eksperimenta i stvarnog ishoda u nadziranom učenju.

		Stvarni ishod	
		Pozitivan	Negativan
Ishod eksperimenta	Pozitivan	Točno pozitivan (TP)	Lažno pozitivan (LP)
	Negativan	Lažno negativan (LN)	Točno negativan (TN)

Standardne metrike vrednovanja najčešće korištene u strojnom učenju su točnost klasifikacije, osjetljivost i specifičnost. One su definirane matricom konfuzije (engl. *confusion matrix*) predstavljenom tablicom 5.1.

Točnost klasifikacije (engl. *classification accuracy*) računa se kao udio ispravno procijenjenih primjeraka u ukupnom broju primjeraka, odnosno: $\frac{TP+TN}{N}$.

Osjetljivost (engl. *sensitivity*) je vjerojatnost uspješnog identificiranja pozitivnih primjeraka. Računa se kao udio ispravno procijenjenih pozitivnih primjeraka u ukupnom broju primjeraka s pozitivnim ishodom, odnosno: $\frac{TP}{TP+LN}$.

Specifičnost (engl. *specificity*) je vjerojatnost uspješnog identificiranja negativnih primjeraka. Računa se kao udio ispravno procijenjenih negativnih primjeraka u ukupnom broju primjeraka s negativnim ishodom, odnosno: $\frac{TN}{TN+LP}$.

Izračunati udjeli u radu se predstavljaju kao postoci. Opisane metrike vrednovanja djelovanja metoda na realnim domenama tretiraju cenzurirane primjerke u podacima za testiranje kao negativne, stoga se moraju interpretirati s određenom dozom opreza. U simulacijskoj je studiji, pak korišten izvorni ishod svakog primjerka, jer je taj bio zabilježen pred postupkom umjetnog cenzuriranja.

Metrike vrednovanja korištene u analizi preživljenja bez problema barataju cenzurom u podacima, stoga se u ovom radu koriste bez ikakvih promjena. Korištene su sljedeće metrike analize preživljenja: indeks suglasnosti, težinska točnost klasifikacije i integrirana Brierova ocjena.

Indeks suglasnosti (engl. *concordance index*) predstavlja vjerojatnost, da za bilo koja dva slučajno odabrana primjerka iz nekog skupa podataka, gdje je onom primjerku s kraćim praćenjem zabilježen pozitivan ishod, upravo taj primjerak ima manju vjerojatnost preživljenja od onog primjerka s dužim praćenjem (Harrell *et al.*, 1982). Ta je vjerojatnost ekvivalentna površini područja ispod ROC krivulje (engl. *receiver operating characteristic curve* (Hanley & McNeil, 1982). Indeks suglasnosti računa se iz podataka kao udio konzistentnih parova primjeraka u ukupnom broju upotrebljivih parova primjeraka. Par primjeraka je upotrebljiv kada je primjerku s kraćim praćenjem zabilježen pozitivan ishod. Par je konzistentan ako je primjerku s kraćim praćenjem predviđena manja vjerojatnost preživljenja nego duže praćenom primjerku.

Težinska točnost klasifikacije (engl. *weighted classification accuracy*), predložena u Ripley & Ripley (2001), izvrsna je zamjena za standardnu točnost klasifikacije u situacijama u kojima je prisutna neizvjesnost u ishodu cenzuriranih primjeraka. Kako bismo nesmetano mogli koristiti cenzurirane podatke kao dio podataka za testiranje, pomoću Kaplan - Meier procjene (Kaplan & Meier, 1958) moguće je procijeniti vjerojatnost preživljenja P_s pojedinog primjerka do kraja njegovog vremena praćenja. Svaki takav primjerak ulazi u podatke za testiranje s oba moguća ishoda, uz težinske faktore

P_s i $1 - P_s$. Ostali primjerci, oni kojima je zabilježen pozitivan ishod, ulaze u podatke za testiranje s težinskim faktorom 1.

Integrirana Brierova ocjena (engl. *integrated Brier score*) iz svog je izvornog oblika (Brier, 1950) bila prilagođena radu s cenzuriranim podacima o preživljenju (Graf *et al.*, 1999). Uz zadovoljenu pretpostavku da je cenzuriranje u podacima slučajno, Brierova ocjena definirana je izrazom:

$$BS^C(t) = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{S}(t|\mathbf{x}_i)^2 I(\tilde{T}_i \leq t, \delta_i = 1)\hat{K}(\tilde{T}_i)^{-1} + (1 - \hat{S}(t|\mathbf{x}_i))^2 I(\tilde{T}_i > t)\hat{K}(t)^{-1}\Big). \tag{5.1}$$

Uz zadanu populaciju veličine N, primjerak $\boldsymbol{x}_i$ ima vrijeme praćenja $\tilde{T}_i = min(T_i, C_i)$, dok je $\delta_i = I(T_i \leq C_i)$ pokazivačka varijata zbivanja ishoda. U slučaju da je uvjet ispunjen, pokazivačka varijata ima vrijednost jedan; u suprotnom ima vrijednost nula. T_i predstavlja vrijeme preživljenja, dok C_i predstavlja vrijeme praćenja. $\hat{K}(t)$ predstavlja Kaplan - Meier procjenu distribucije cenzuriranja K, baziranu na svim primjercima $(\tilde{T}_i, 1 - \delta_i)$. Doprinos svakog primjerka ukupnoj ocjeni u trenutku t jednak je kvadratu razlike između zabilježenog ishoda u trenutku t i predviđane vjerojatnosti preživljenja u t, normaliziranom kako bi se nadoknadio gubitak informacije zbog utjecaja cenzure. Pod pretpostavkom da je cenzuriranje slučajno, integrirana Brierova ocjena (IBS) određena je izrazom:

$$IBS^C = max(\tilde{T}_i)^{-1}\int_0^{max(\tilde{T}_i)} BS^C(t)dt. \tag{5.2}$$

Mjera objašnjene rezidualne varijacije (engl. *residual variation*, RV) pokazuje relativno poboljšanje integrirane Brierove ocjene testiranog prediktora u odnosu na neparametarsko Kaplan - Meier predviđanje IBS_0^C:

$$R^2 = 1 - \frac{IBS^C}{IBS_0^C}. \tag{5.3}$$

Ukoliko je rezidualna varijacija R^2 pozitivna, prediktor je točniji od "naivnog" Kaplan - Meier predviđanja (ono je za sve primjerke jednako, to jest ne ovisi o vrijednostima poznatih kovarijata). U Graf *et al.* (1999) se, kao gornja vremenska granica za izračun integrirane Brierove ocjene, predlaže korištenje medijane svih vremena praćenja umjesto najdužeg vremena praćenja. Razlozi za takvu intervenciju nalaze se u činjenici da prediktori s vremenom postaju sve manje točni, zbog većeg utjecaja cenzure.

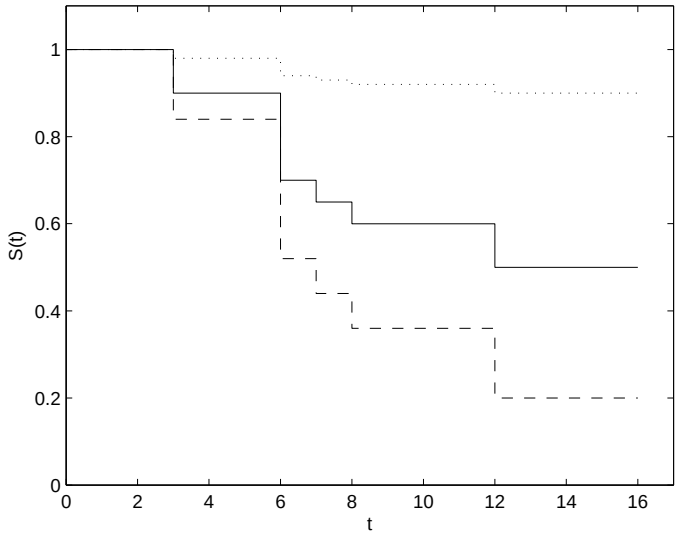
5.2 Postupci vrednovanja

Kako bi se kod naučenih modela vrednovala sposobnost generalizacije, odnosno kako bi se u testiranju i interpretaciji rezultata spriječila pristranost zbog pretreniranosti modela

podacima (engl. *overfitting*), u ovom se radu koristi vrednovanje podjelom na podatke za učenje i podatke za testiranje (engl. *learn-test split methodology*) (Witten & Frank, 2005). Kada ne postoji odvojeni skup podataka za testiranje, može se (slučajno) izdvojiti iz izvornog skupa podataka. U ovom je slučaju korišten postupak iterativne stratificirane n-struke unakrsne provjere (engl. *iterative stratified n-fold cross-validation*) (Michie *et al.*, 1995). Taj postupak prvo slučajno podijeli skup primjeraka na n disjunktivnih podskupova približno jednakih veličina (broja primjeraka) i približno jednake distribucije među različitim ishodima. Zatim se u odvojenim pokusima uči model iz $n-1$ podskupova podataka te se testira na preostalom podskupu podataka. Konačan rezultat vrednovanja računa se kao aritmetička sredina dobivenih vrijednosti u svakom od n pokusa, na nekom broju iteracija. Srednje su vrijednosti dobivene testovima popraćene standardnim devijacijama te su u svim tablicama predstavljene kao $\bar{x}(s)$. Podaci su u simulacijskim studijama i provjeri na realnim domenama dobrim dijelom stratificirani "na slijepo", jer je za određivanje podjela po kriteriju približno jednakih distribucija među različitim ishodima bio korišten cenzuriran ishod. Radi jednostavnosti, u svim je pokusima korištena deseterostruka unakrsna provjera, osim u statističkim testovima nad realnim domenama (sekcija 5.2.1).

Za usporedbu sličnosti struktura Bayesovih mreža, naučenih iz podataka, strukturama ishodišnih Bayesovih mreža (onih iz kojih su podaci za učenje uzorkovani; za detalje vidi sekciju 6.2.1), postupak unakrsne provjere nije bio potreban, stoga nije bio ni korišten. Razlog tome leži u činjenici da su se naučeni modeli vrednovali, ne na podacima za učenje, već na topologijama ishodišnih modela. Sličnosti i razlike u strukturama izmjerene su računanjem broja naučenim modelima dodanih (suvišnih), oduzetih (nedostajućih) i obrnutih (preokrenutih) lukova, u odnosu na ishodišne modele. Izmjerene vrijednosti su naknadno pretvorene u postotke ukupnog broja lukova radi preglednijeg predstavljanja. Ovaj se postupak, nažalost, nije dao primijeniti u radu s realnim domenama, jer su njihovi ishodišni (inherentni) modeli nepoznati.

Računanje integrirane Brierove ocjene i njene rezidualne varijacije, predviđeno radu sa statističkim postupcima za analizu preživljenja, prilagođeno je modelima dobivenim korištenim postupcima strojnog učenja i njihovim derivatima: algoritmu penjanja uzbrdo, algoritmu uvjetnih neovisnosti i naivnom Bayesovom klasifikatoru. Za razliku od modela proporcionalnih hazarda, koji na osnovi vrijednosti poznatih kovarijata pronalazi funkciju vjerojatnosti preživljenja (neovisna varijata je vrijeme), ostali modeli predviđaju jednu jedinu vjerojatnost, koju se tretira kao stvarnu i konačnu vjerojatnost preživljenja. Pošto je vjerojatnije da će dani primjerak preživjeti kraće, nego da će preživjeti duže, ta je konačna vjerojatnost preživljenja bila projicirana na krivulju preživljenja (pretpostavljeno je da je cenzuriranje slučajno). Konačna vjerojatnost preživljenja odnosi se na vjerojatnost preživljenja s najdužim vremenom praćenja t^*. Za projiciranje je korišteno preslikavanje na Kaplan - Meier procjenu funkcije preživljenja $\hat{S}_0(t)$, kao što je predstavljeno na slici 5.1. Ukoliko je predviđena vjerojatnost preživljenja viša od konačne Kaplan - Meier procjene vjerojatnosti preživljenja $\hat{S}_0(t^*)$, krivulja preživljenja $S(t)$ se diže (razlomljena linija označena točkicama); ukoliko je niža, krivulja se spušta (razlomljena linija označena crticama). Projicirana predviđena funkcija



Slika 5.1: Primjer projekcije konačne vjerojatnosti preživljenja P_S na krivulju preživljenja $S(t)$. Puna linija predstavlja Kaplan - Meier procjenu krivulje preživljenja za skup primjeraka; linija označena crticama predstavlja projekciju krivulje preživljenja za $P_S = 0.9$; linija označena točkicama predstavlja projekciju krivulje preživljenja za $P_S = 0.2$.

vjerojatnosti računa se izrazom:

$$S(t) = 1 - \left(1 - \hat{S}_0(t)\right) \cdot \frac{1 - P_S}{1 - \hat{S}_0(t^*)}. \tag{5.4}$$

Kao primjer pogledajmo izračun vjerojatnosti preživljenja u trenutku $t = 10$, $S(10)$, za predviđanu konačnu vjerojatnost preživljenja $P_S = 0.2$. Budući da $\hat{S}_0(t^*) = 0.5$ i $\hat{S}_0(10) = 0.6$ (slika 5.1), dobijemo $S(10) = 1 - (1 - 0.6) \cdot (1 - 0.2)/(1 - 0.5) = 0.36$.

5.2.1 Statističko vrednovanje

Kako bi se dobila procjena stvarne učinkovitosti opisanih postupaka prilagodbe podataka o preživljenju za algoritme strojnog učenja, nad dobivenim je rezultatima potrebno izvesti statističke testove (Demšar, 2006). Prikladan neparametarski test za usporedbu više postupaka učenja na više skupova podataka je Friedmanov test (Friedman, 1937). Friedmanov test rangira rezultate vrednovanja postupaka za svaki skup podataka posebno, to jest dodijeli 1. mjesto najboljem, 2. mjesto idućem najboljem itd. U slučaju da dva postupka imaju jednako dobar rezultat, dodijeli im se srednji rang (npr. ako su 3. i 4. jednaki, obama se dodijeli rang 3.5).

Neka je g_i^j rang j-tog od k postupaka na i-tom od m skupova podataka. Friedmanov

test uspoređuje prosječne rangove postupaka, $G_j = \frac{1}{m}\sum_i g_i^j$. Pod nultom hipotezom, koja određuje da su svi postupci ekvivalentni, što bi značilo da bi i njihovi prosječni rangovi trebali biti jednaki, Friedmanova statistika:

$$\chi_F^2 = \frac{12m}{k(k+1)} \cdot \left[\sum_j G_j^2 - \frac{k(k+1)^2}{4} \right], \tag{5.5}$$

distribuirana je po χ^2 s $k-1$ stupnjeva slobode, kada su m i k dovoljno veliki (Demšar, 2006). Iman & Davenport (1980) predlažu bolju statistiku, temeljenu na Friedmanovoj:

$$F_F = \frac{(m-1)\chi_F^2}{m(k-1)-\chi_F^2}. \tag{5.6}$$

Ta je distribuirana po F distribuciji s $k-1$ i $(k-1)\cdot(m-1)$ stupnjeva slobode. Ukoliko je nulta hipoteza eksperimentalno odbačena, postupci nisu jednako učinkoviti te je potrebno usporediti svaki sa svim ostalim Nemenyijevim testom (Nemenyi, 1963). Dva se postupka razlikuju po učinkovitosti ako je razlika njihovih srednjih rangova veća ili jednaka vrijednosti kritične razlike:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6m}}, \tag{5.7}$$

gdje su kritične vrijednosti q_α temeljene na studentiziranoj statistici ranga podijeljenoj s $\sqrt{2}$ (Demšar, 2006).

Usporedba više postupaka učenja na više skupova podataka korištena je u simulacijskoj studiji (poglavlje 6), jer je broj međusobno nezavisnih skupova podataka u njoj bio dovoljno velik. Rezultati usporedbe predstavljeni su u obliku grafova srednjih rangova, predloženim u Demšar (2006). Kod realnih je domena (poglavlje 7) taj broj premalen, zbog čega su postupci učenja uspoređeni na svakoj od realnih domena zasebno korištenjem neparametarskog Friedmanovog dvostranog ANOVA testa rangiranjem (Sheskin, 2004). Test se vrši na rezultatima dobivenim u 5 iteracija dvodijelne unakrsne provjere, po uzoru na združeni F test unakrsne provjere 5 x 2 (engl. *combined 5x2 cv F test*) (Alpaydin, 1999). Ukoliko je nulta hipoteza eksperimentalno odbačena, nisu svi postupci jednako učinkoviti, te je potrebno usporediti svakog sa svakim, korištenjem neparametarskog Wilcoxonovog dvostranog testa rangiranih predznaka (engl. *Wilcoxon signed-rank test*), sa statističkom razinom značajnosti testa $\alpha = 0.05$. Razina značajnosti testa ispravljena je Bonferronijevom korekcijom, koja je nužna zbog relativnog povećanja pogreške istovremenim testiranjem više hipoteza. Za izvođenje Friedmanovog dvostranog ANOVA testa rangiranjem i *post-hoc* testova korištene su gotove funkcije okruženja MATLAB (sekcija 1.2).

5.3 Oznake postupaka

U ovoj je disertaciji uspoređen učinak korištenja raznih postupaka prilagodbe podataka o preživljenju (poglavlje 3) za razne postupke strojnog učenja (poglavlje 4). Na nekim

je mjestima bilo potrebno, zbog manjka prostora, koristiti skraćene nazive kombinacija tih postupaka, npr. na slikama i u tablicama. Skraćeni nazivi su objašnjeni u nastavku.

Postupak učenja Bayesove mreže algoritmom penjanja uzbrdo označen je s HC, dok je za algoritam uvjetnih neovisnosti označen s CI. Postupak učenja naivnog Bayesovog klasifikatora označen je s NB, a regresijskog modela proporcionalnih hazarda s COX. Postupci prilagodbe podataka o preživljenju označavaju se dodavanjem prefiksa ili sufiksa na navedene oznake postupaka strojnog učenja (HC, CI, NB): postupak tretiranja cenzuriranih primjeraka kao negativnih nema dodatne oznake (HC, CI, NB); postupak odstranjivanja prekratko praćenih primjeraka označen je prefiksom c (cHC, cCI, cNB); postupak podjele na vremenske intervale označen je prefiksom i (iHC, iCI, iNB); postupak podvajanja cenzuriranih primjeraka uz težinske faktore označen je prefiksom w (wHC, wCI, wNB); postupak podvajanja cenzuriranih primjeraka uz težinske faktore isključivo za treniranje parametara Bayesove mreže (mreža je naučena tretiranjem cenzuriranih primjeraka kao negativnih) označena je prefiksom s (sHC, sCI); postupak odstranjivanja šuma cenzure označen je sufiksom f (HCf, CIf, NBf). U radu s realnim domenama bilo je potrebno predstaviti i rezultate učenja iz nediskretiziranih podataka za regresijski model proporcionalnih hazarda - tu je dodan sufiks c (COXc).

Poglavlje 6

Vrednovanje simulacijskom studijom

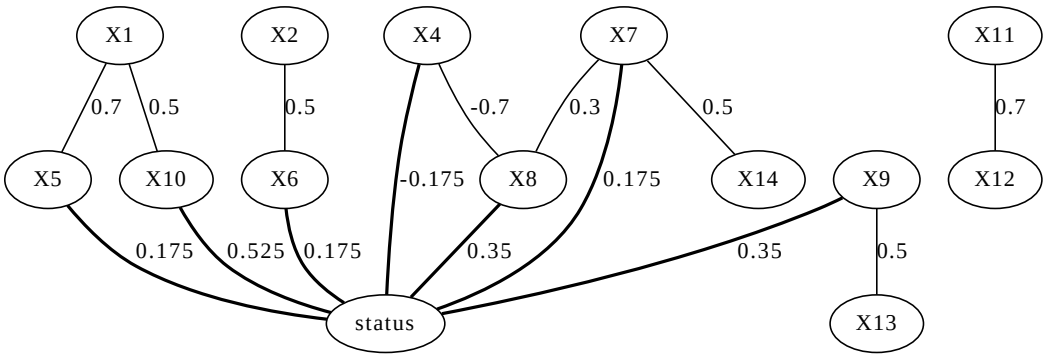
U ovom su poglavlju predstavljene simulacijske studije utjecaja cenzure na korištene algoritme strojnog učenja, uz različite postupke prilagodbe podataka o preživljenju. U prvoj simulacijskoj studiji (6.1) ispitana je učinkovitost navedenih postupaka da nauče modele sposobne za ispravnu klasifikaciju njima nepoznatih primjeraka. U drugoj je studiji (6.2) ispitana učinkovitost navedenih postupaka u učenju ispravnih topologija Bayesovih mreža. Obje bi studije trebale čitatelju dati jasnu sliku o karakteristikama i mogućnostima korištenih postupaka.

6.1 Studija učinkovitosti modela

Poradi jednostavnosti, razumljivosti te univerzalne upotrebljivosti u ovoj je simulacijskoj studiji korišten postupak opisan u Royston & Sauerbrei (2004). Postupak je djelomično izmijenjen radi prilagodbe potrebama ove studije, čiji je cilj simuliranje tipičnog obrasca učinaka često viđenog u studijama prognostičkih faktora. U nastavku je detaljno predstavljen ishodišni model (6.1.1) te postupak generiranja (6.1.2) i cenzuriranja (6.1.3) primjeraka, čemu slijedi predstavljanje i interpretacija rezultata (6.1.4).

6.1.1 Ishodišni model

Korišteno je 15 kovarijata $\boldsymbol{X} = (X_1, \ldots, X_{15})'$. Svaka od njih ima multivarijatnu normalnu distribuciju sa srednjom vrijednošću 0, varijancom 1 i korelacijama $\rho_{i,j} = corr(X_i, X_j) = 0$, osim za $\rho_{1,5} = 0.7$, $\rho_{1,10} = 0.5$, $\rho_{2,6} = 0.5$, $\rho_{4,8} = -0.7$, $\rho_{7,8} = 0.3$, $\rho_{7,14} = 0.5$, $\rho_{9,13} = 0.5$ i $\rho_{11,12} = 0.7$. Vektor regresijskih koeficijenata je određen kao $\boldsymbol{\beta'} = (0, 0, 0, -0.175, 0.175, 0.175, 0.175, 0.35, 0.35, 0.525, 0, 0, 0, 0, 0)$. Iz opisa vektora da se vidjeti kako osam kovarijata ne utječe na ishod, dočim sedam njih utječe. Jedan od tih drugih (β_{10}), značajno je veći od ostalih. Interakcija varijata predstavljena je na slici 6.1.



Slika 6.1: Korelacijski (uz normalnu crtu) i regresijski (uz debelo otisnutu crtu) koeficijenti, predstavljeni u obliku grafa.

6.1.2 Generiranje primjeraka

Iz opisane korelacijske matrice generira se niz primjeraka. Za potrebe simulacijske studije bilo je nužno generirati i binarni ishod za svaki primjerak. To je napravljeno korištenjem logističke regresije. Logistička regresija je generalizacija linearne regresije (Hastie *et al.*, 2001), koja se najčešće koristi za predviđanje binarnih ovisnih varijata. Vjerojatnost da je generirani ishod primjerka $\boldsymbol{x}$ pozitivan računa se izrazom:

$$P(O^+) = \frac{e^{\boldsymbol{\beta' x}}}{1 + e^{\boldsymbol{\beta' x}}}. \tag{6.1}$$

Primjercima s izračunatom vjerojatnošću pozitivnog ishoda većom od 50% dodijeljen je pozitivan ishod, to jest pretpostavljeno je da se za takve primjerke događaj od interesa zbio, dok je ostalima dodijeljen negativan ishod, to jest pretpostavljeno je da se za takve primjerke događaj od interesa nije zbio. Ishod je jednakomjerno distribuiran po oba slučaja.

Većina korištenih postupaka nije u mogućnosti učiti iz numeričkih (kontinuiranih) podataka; iz tog su razloga sve numeričke varijate diskretizirane postupkom podjele na jednake intervale (engl. *equal-width binning*). Isti diskretizirani podaci korišteni su i u postupcima koji mogu rukovati numeričkim podacima, kako bi svi bili podvrgnuti jednakim uvjetima testiranja. Postoje prikladniji postupci diskretizacije cenzuriranih podataka, poput najboljeg log-rank razdvajanja (Contal & O'Quigley, 1999; Klein & Moeschberger, 2003), no ti namjerno nisu korišteni jer bi pojednostavili problem učenja izbacivanjem za ishod nebitnih numeričkih kovarijata. S obzirom na to da je u ovoj simulacijskoj studiji bilo moguće generirati proizvoljan, to jest dovoljno velik broj primjeraka, na taj se način mogla provjeriti sposobnost postupaka strojnog učenja da sami pronađu kovarijate korelirane s ishodom, te da pritom one nebitne zanemare.

6.1.3 Cenzuriranje primjeraka

Cenzuriranje se pojavljuje kada je pacijent iz studije povučen prije nego se ishod stigao zbiti. U ovom se umjetnom testu cenzura može ustanoviti usporedbom vremena preživljenja i vremena promatranja. Vrijeme preživljenja se određuje korištenjem eksponencijalne distribucije (Bender *et al.*, 2005) kovarijata i regresijskih koeficijenata, uz pretpostavku da je $\lambda_T = 0.002$. Vrijeme preživljenja određeno je izrazom:

$$T_i = -\frac{\ln \upsilon_i}{\lambda_T \cdot e^{\boldsymbol{\beta}' \boldsymbol{x_i}}}, \tag{6.2}$$

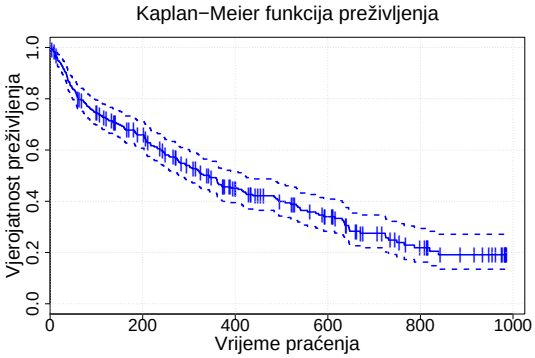
u kojem je υ_i uzorkovan iz pseudoslučajne uniformne distribucije $U(0, 1)$. Za svako generirano vrijeme preživljenja, na isti je način generirano dodatnih 8 pseudoslučajno eksponencijalno distribuiranih vremena promatranja za 8 različitih razina cenzure, s hazardima $\lambda_C = 0.0003$ (10%-tna cenzura), 0.0004 (20%-tna cenzura), 0.0006 (30%-tna cenzura), 0.0012 (40%-tna cenzura), 0.002 (50%-tna cenzura), 0.0033 (60%-tna cenzura), 0.0067 (70%-tna cenzura) i 0.01 (80%-tna cenzura). Generirano vrijeme preživljenja T_i se u svakoj od postava cenzure (od 10% do 80%) smatra cenzuriranim ako je veće od pripadajućeg generiranog vremena promatranja u toj postavi (Royston & Sauerbrei, 2004). Navedeni postoci cenzure odnose se na udio cenzuriranih primjeraka u ukupnom broju primjeraka s pozitivnim ishodom. Funkcije preživljenja generiranih podataka po svim postavama cenzure predstavljene su na slici 6.2 (odnose se na prvi od skupova generiranih podataka). Vrijeme opažanja na slici ograničeno je na 1000 (virtualnih vremenskih jedinica), kako bi se naglasile razlike među pojedinim funkcijama preživljenja.

Postupak generiranja i cenzuriranja primjeraka po postavama cenzure izveden je ukupno 100 puta. Rezultati opisani u nastavku predstavljaju srednje vrijednosti i standardne devijacije za tih 100 umjetno generiranih skupova podataka. Svaki generirani skup podataka sadrži točno 400 primjeraka. Eksperimentalno je utvrđeno da se povećanjem broja primjeraka po skupu podataka rezultati vrednovanja u ovoj simulacijskoj studiji nisu dodatno popravili.

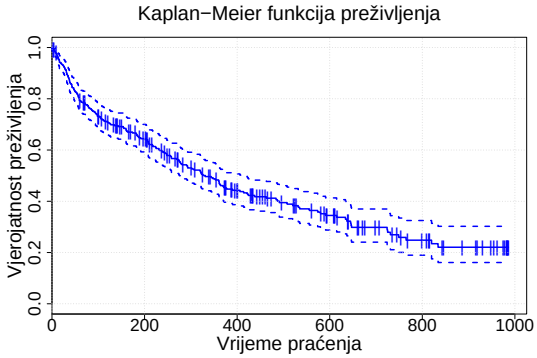
6.1.4 Rezultati

Odabrane granice za postupak prilagodbe podataka podjelom na vremenske intervale su $(0, 300, 800, \infty)$. Rezultati pokusa ove simulacijske studije standardnim metrikama predstavljeni su slikama 6.3 (točnost klasifikacije), 6.4 (osjetljivost) i 6.5 (specifičnost). Oznake korištenih postupaka opisane su u sekciji 5.3. Rezultati upućuju na očiglednu superiornost postupka prilagodbe podataka odstranjivanjem šuma cenzure (HCf, CIf i NBf), posebno pri visokom udjelu cenzure u podacima (više od 50%). Za srednju razinu cenzure (od 20% do 50%) najbolji je postupak prilagodbe podataka podvajanjem cenzuriranih primjeraka uz težinske faktore (wHC, wCI i wNB). Modeli dobiveni učenjem i strukture Bayesovih mreža i njihovih parametara iz takvih podvojenih podataka (wHC i wCI), po točnosti klasifikacije ne razlikuju se od onih u kojima se struktura učila tretiranjem cenzuriranih primjeraka kao negativnih (sHC i sCI). Uz nisku razinu cenzure svi postupci prilagodbe podataka za učenje Bayesovih mreža, osim postupka odstranjiva-

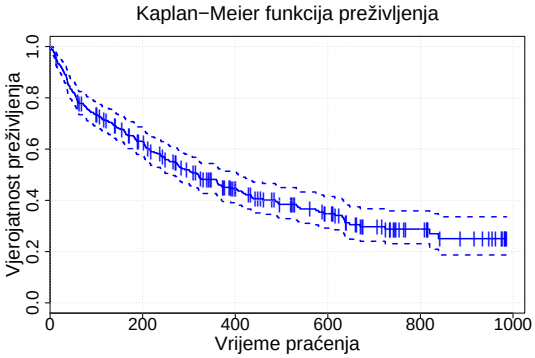
nja prekratko praćenih primjeraka, po točnosti klasifikacije nadmašuju Coxovu regresiju (COX). Postupak odstranjivanja prekratko praćenih primjeraka ima očekivano lošije rezultate od ostalih postupaka, jer je cenzura u podacima slučajna. Tim je postupkom efektivno slučajno odstranjena polovica primjera za učenje, što otežava posao algoritmima za učenje Bayesovih mreža. Isti postupak prilagodbe podataka učenju naivnog Bayesovog klasifikatora manje škodi, jer taj očito ne treba toliko puno podataka za ko-



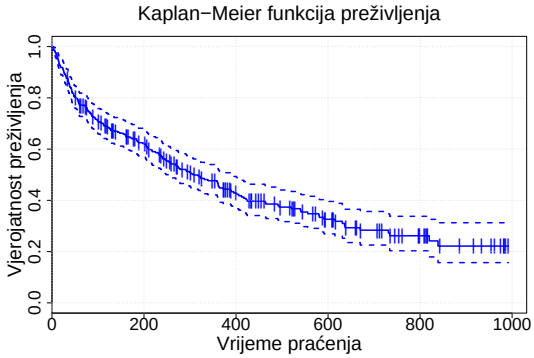
(a) Bez cenzure



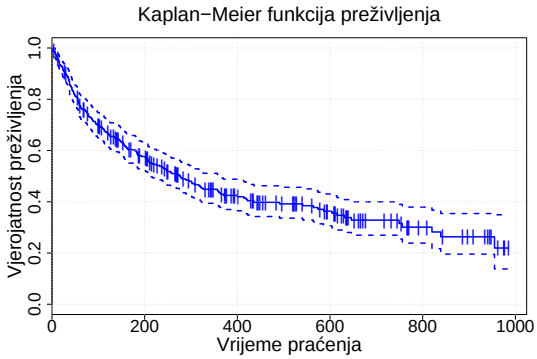
(b) Cenzura 10%



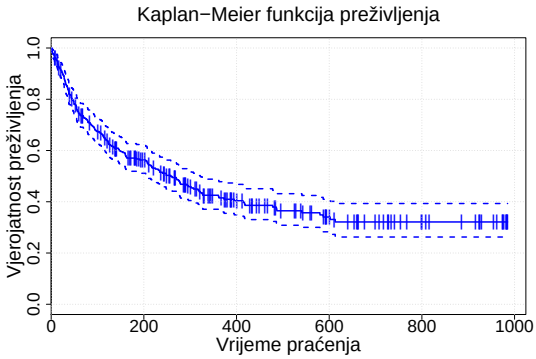
(c) Cenzura 20%



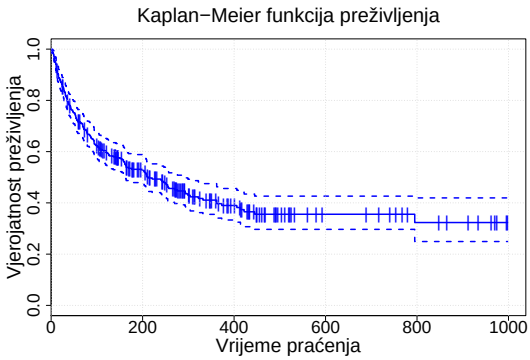
(d) Cenzura 30%



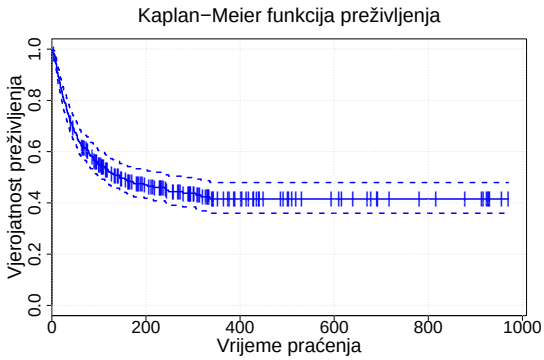
(e) Cenzura 40%



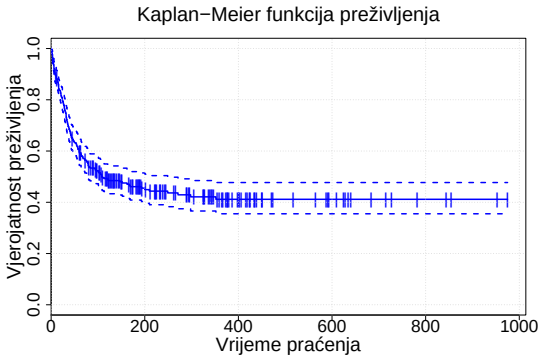
(f) Cenzura 50%



(g) Cenzura 60%



(h) Cenzura 70%

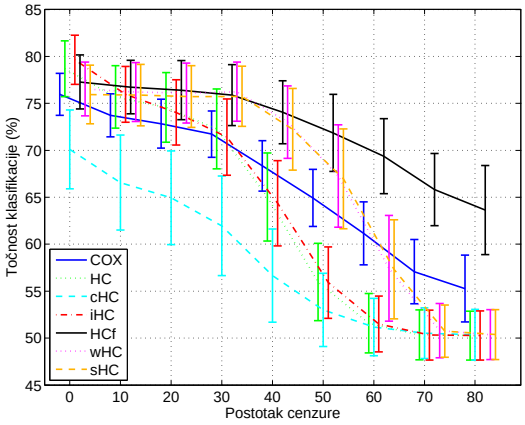


(i) Cenzura 80%

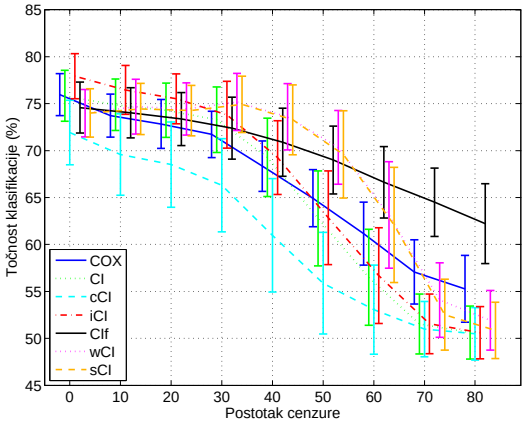
Slika 6.2: Funkcije preživljenja za svaku od postava cenzure u simulacijskoj studiji. Funkcije su praćene krivuljama 95%-tnih intervala pouzdanosti (crtkano).

nvergiranje k smislenom rješenju. Postupak podjele podataka na vremenske intervale se po točnosti klasifikacije gotovo nimalo ne razlikuje od postupka tretiranja cenzuriranih primjeraka kao negativnih. Zamjetna je razlika u osjetljivosti kod postupka prilagodbe podataka odstranjivanjem šuma cenzure, koja je u odnosu na ostale postupke očigledno viša, i specifičnosti, koja je očigledno niža. To je posljedica povećanja udjela pozitivnih primjeraka u podacima s takvim postupkom prilagodbe podataka.

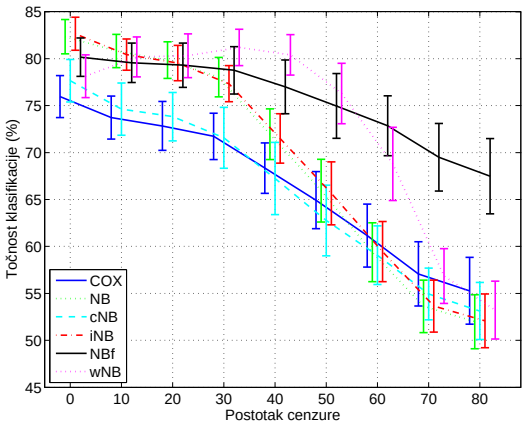
Rezultati na metrikama analize preživljenja predstavljeni su slikama 6.6 (težinska točnost klasifikacije), 6.7 (indeks suglasnosti), 6.8 (integrirana Brierova ocjena) i 6.9 (rezidualna varijacija integrirane Brierove ocjene). Uspoređivanjem krivulja sa slike 6.3 i slike 6.6, vidljivo je da je težinska točnost klasifikacije izvrsna zamjena za standardnu točnost klasifikacije (relativan odnos krivulja je sličan). Ona pravilno potvrđuje relativan odnos različitih postupaka prilagodbe podataka po modelima, određen mjerom točnosti klasifikacije. Ostale mjere analize preživljenja preferiraju Coxovu regresiju. Postupak odstranjivanja šuma cenzure kod algoritama za učenje Bayesovih mreža (HCf, CIf) ima viši indeks suglasnosti od ostalih postupaka prilagodbe podataka samo uz visok udio cenzure (70% i 80%). Integrirana Brierova ocjena (i njena rezidualna varijacija)



(a) Algoritam penjanja uzbrdo

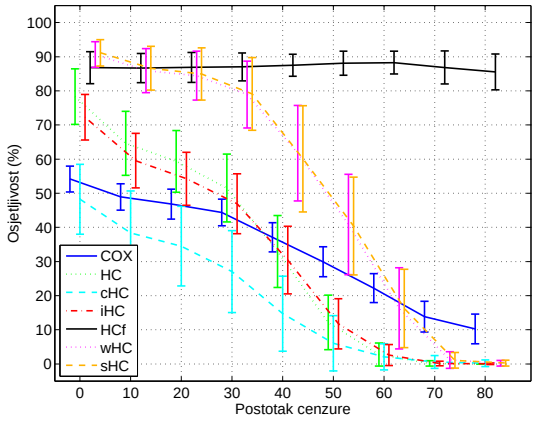


(b) Algoritam uvjetnih neovisnosti

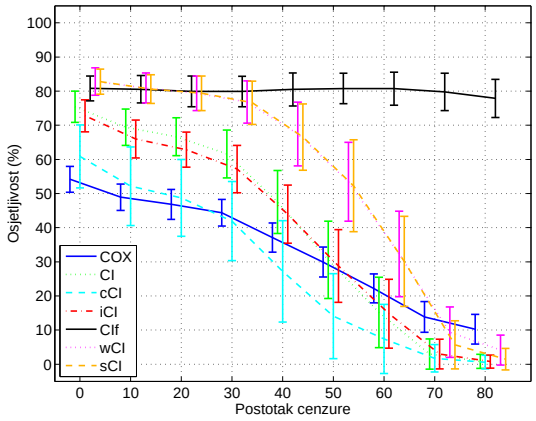


(c) Naivni Bayesov klasifikator

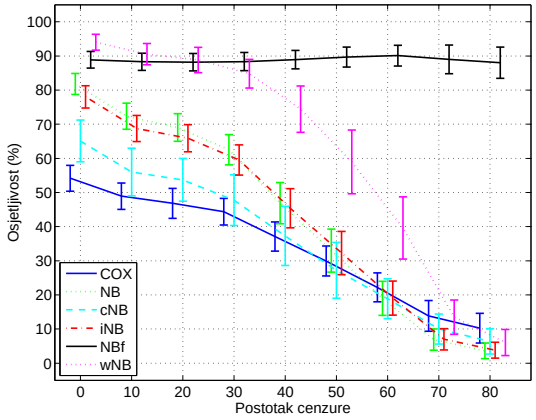
Slika 6.3: Točnost klasifikacije $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



(a) Algoritam penjanja uzbrdo

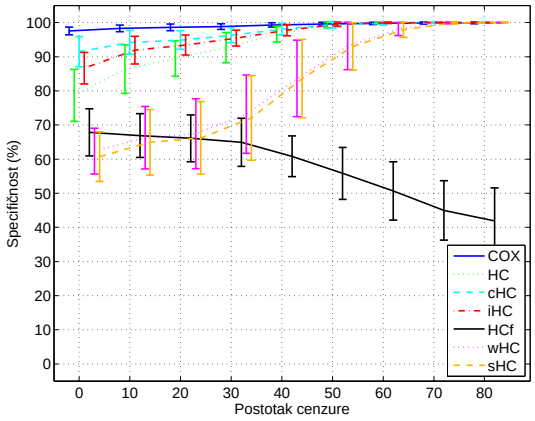


(b) Algoritam uvjetnih neovisnosti

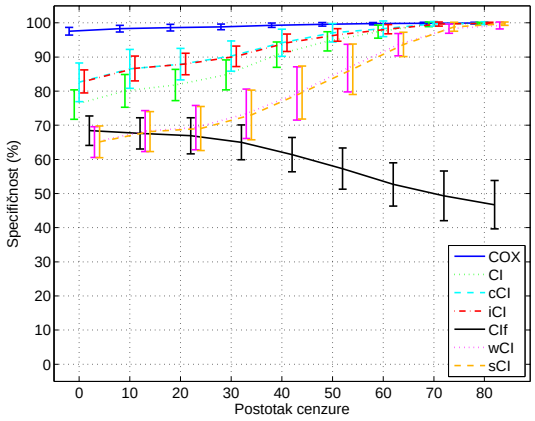


(c) Naivni Bayesov klasifikator

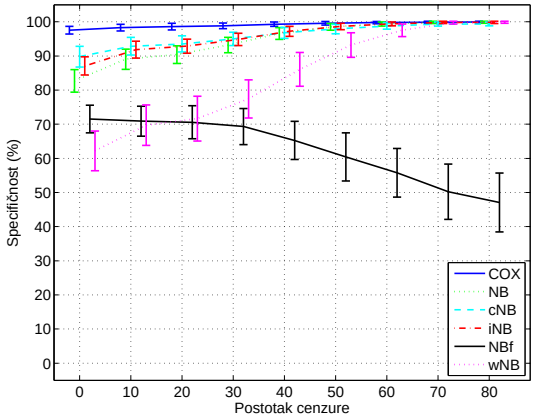
Slika 6.4: Osjetljivost $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



(a) Algoritam penjanja uzbrdo



(b) Algoritam uvjetnih neovisnosti



(c) Naivni Bayesov klasifikator

Slika 6.5: Specifičnost $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.

kod algoritama za učenje Bayesovih mreža ovo potvrđuje. Uz to dodatno ispravlja procjenu dobivenu indeksom suglasnosti kod naivnog Bayesovog klasifikatora (NBf). Od postupaka prilagodbe podataka uz nižu i srednju razinu cenzure (do 50%), indeks suglasnosti kao najbolji ističe postupak podjele na vremenske intervale (iHC, iCI, iNB). Sve metrike analize preživljenja potvrđuju da je najlošiji postupak prilagodbe podataka odstranjivanjem prekratko praćenih primjeraka.

Na slikama 6.10, 6.11 i 6.12 prikazani su srednji rangovi točnosti klasifikacije postupaka učenja (penjanje uzbrdo, uvjetne neovisnosti, naivni Bayesov klasifikator) s različitim postupcima prilagodbe podataka za sve razine cenzure (od 0% do 80%). Za svaku razinu cenzure prikazani su srednji rangovi različitih postupaka prilagodbe podataka i Coxove regresije, što je usporedivo s rezultatima točnosti klasifikacije na slici 6.3. Postupci koji su po performansama sličniji, na linijama se nalaze bliže jedni drugima. Statistički su slični oni postupci čija je razlika srednjih rangova manja od vrijednosti kritične razlike (sekcija 5.2.1). Statistički slični postupci na nekoj razini cenzure povezani su crtom; oni koji nisu povezani crtom, značajno su različiti na statističkoj razini značajnosti testa $\alpha = 0.05$. Evidentan je rast performansi Coxove regresije i postupka prilagodbe podataka odstranjivanjem šuma cenzure (za sve modele) s rastom udjela cenzure u podacima. Prilagodba podataka odstranjivanjem šuma cenzure statistički je najbolja od 50% udjela cenzure za algoritam penjanja uzbrdo, od 70% za algoritam uvjetnih neovisnosti te od 60% za naivni Bayesov klasifikator. Kod srednjeg udjela cenzure (približno 30% do 50%) uglavnom je najbolji postupak prilagodbe podataka podvajanjem cenzuriranih primjeraka uz težinske faktore. Slike koje prikazuju statističku usporedbu srednjih rangova ostalih metrika su u dodatku C.

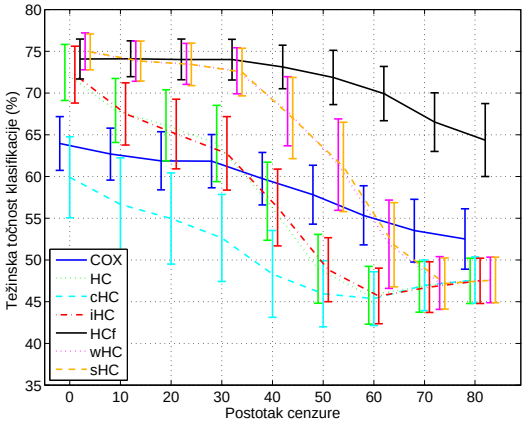
6.2 Studija otkrivanja topologija mreža

U nastavku je predstavljena studija koja vrednuje sposobnost algoritama da ispravno nauče strukture Bayesovih mreža, primjenom raznih postupaka prilagodbe podataka o preživljenju. Ova je studija ograničena isključivo na postupke vezane uz učenje Bayesovih mreža iz podataka; tim su povodom iz nje izostavljeni postupci vezani uz naivni Bayesov klasifikator i uz metodu proporcionalnih hazarda.

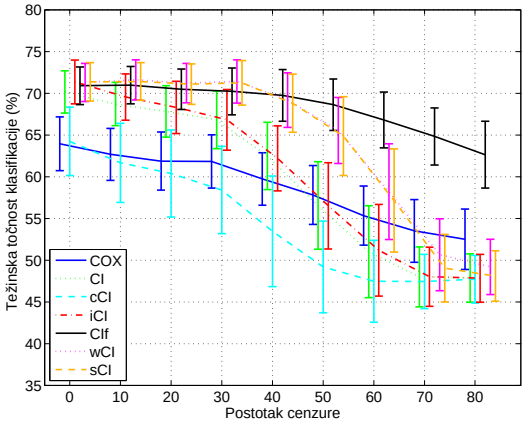
Ishodišni model je u ovoj studiji umjetno generirana Bayesova mreža (sekcija 6.2.1). Topologija mreže odredi se slučajnim dodavanjem usmjerenih lukova praznom grafu uz neka ograničenja. Nakon određivanja topologije, svakom se čvorištu mreže dodijeli tablica (uvjetnih) distribucija vjerojatnosti. Iz generirane se mreže uzorkuju i naknadno umjetno cenzuriraju primjerci (sekcija 6.2.2). Postupci za učenje Bayesovih mreža se zatim vrednuju usporedbom ishodišnih modela i modela naučenih iz uzorkovanih primjeraka (sekcija 6.2.3).

6.2.1 Generiranje ishodišnih Bayesovih mreža

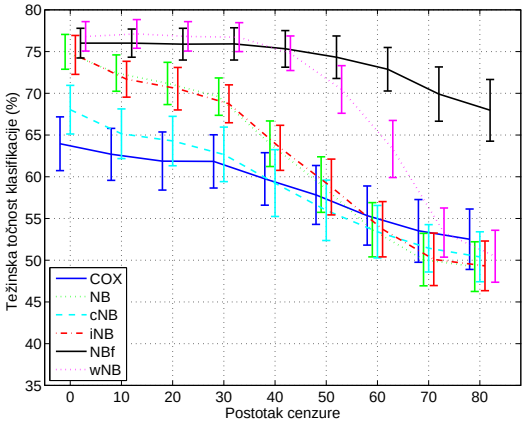
Ishodišna Bayesova mreža sastavljena je od 15 čvorišta ($V_1, \ldots, V_o, \ldots, V_{15}$), od kojih svako čvorište, osim V_o koje predstavlja ishod odnosno varijatu od interesa, predstavlja



(a) Algoritam penjanja uzbrdo

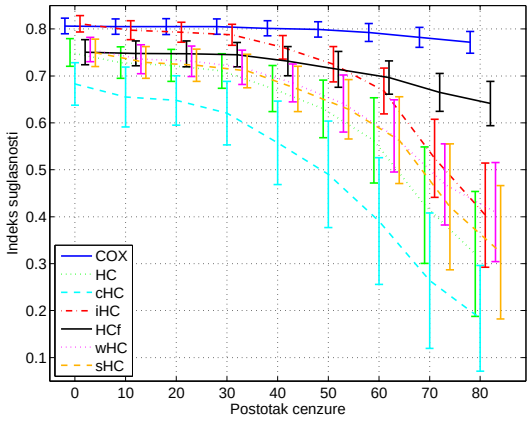


(b) Algoritam uvjetnih neovisnosti

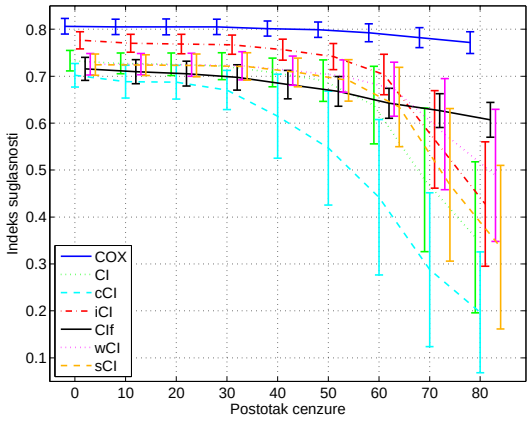


(c) Naivni Bayesov klasifikator

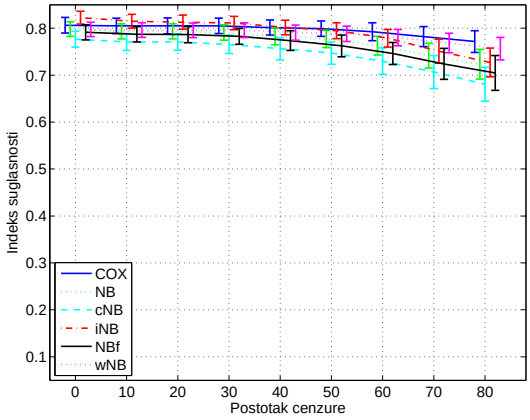
Slika 6.6: Težinska točnost klasifikacije $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



(a) Algoritam penjanja uzbrdo

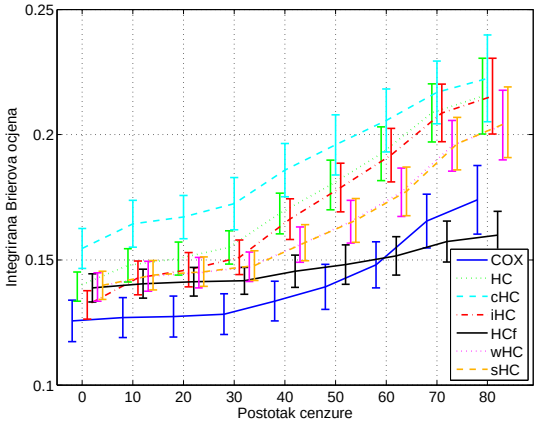


(b) Algoritam uvjetnih neovisnosti

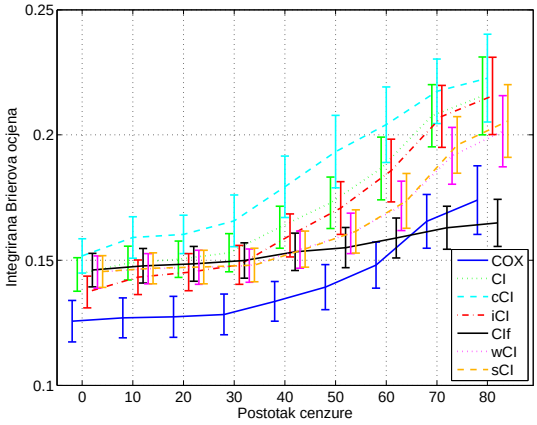


(c) Naivni Bayesov klasifikator

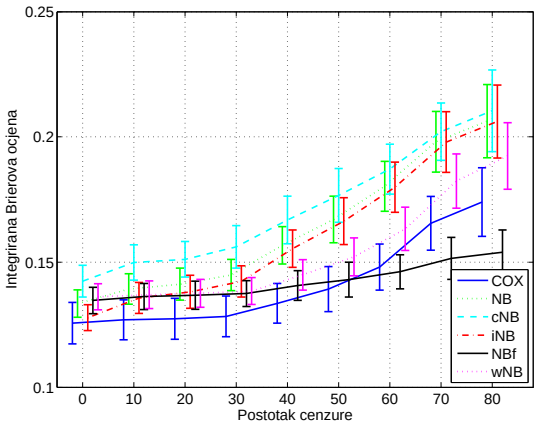
Slika 6.7: Indeks suglasnosti $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



(a) Algoritam penjanja uzbrdo

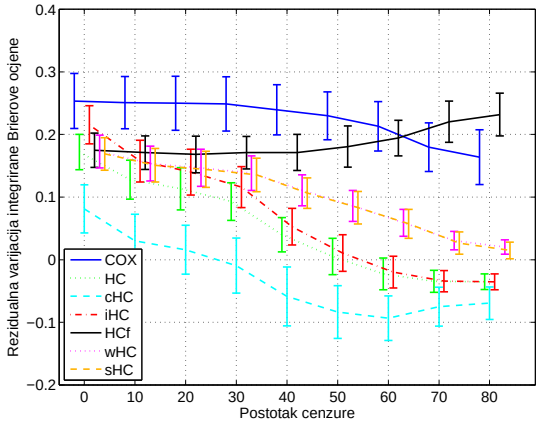


(b) Algoritam uvjetnih neovisnosti

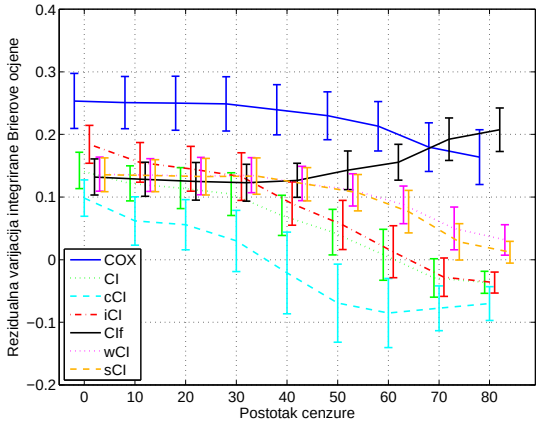


(c) Naivni Bayesov klasifikator

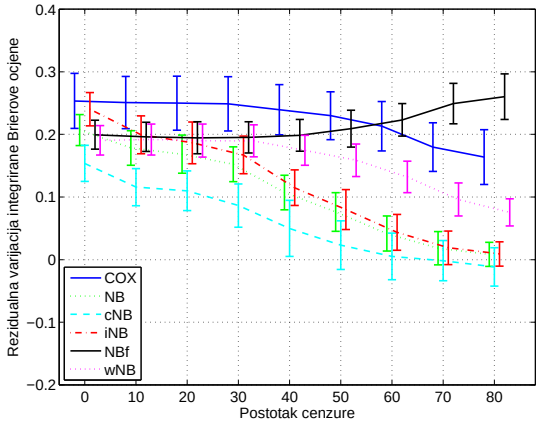
Slika 6.8: Integrirana Brierova ocjena $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



(a) Algoritam penjanja uzbrdo

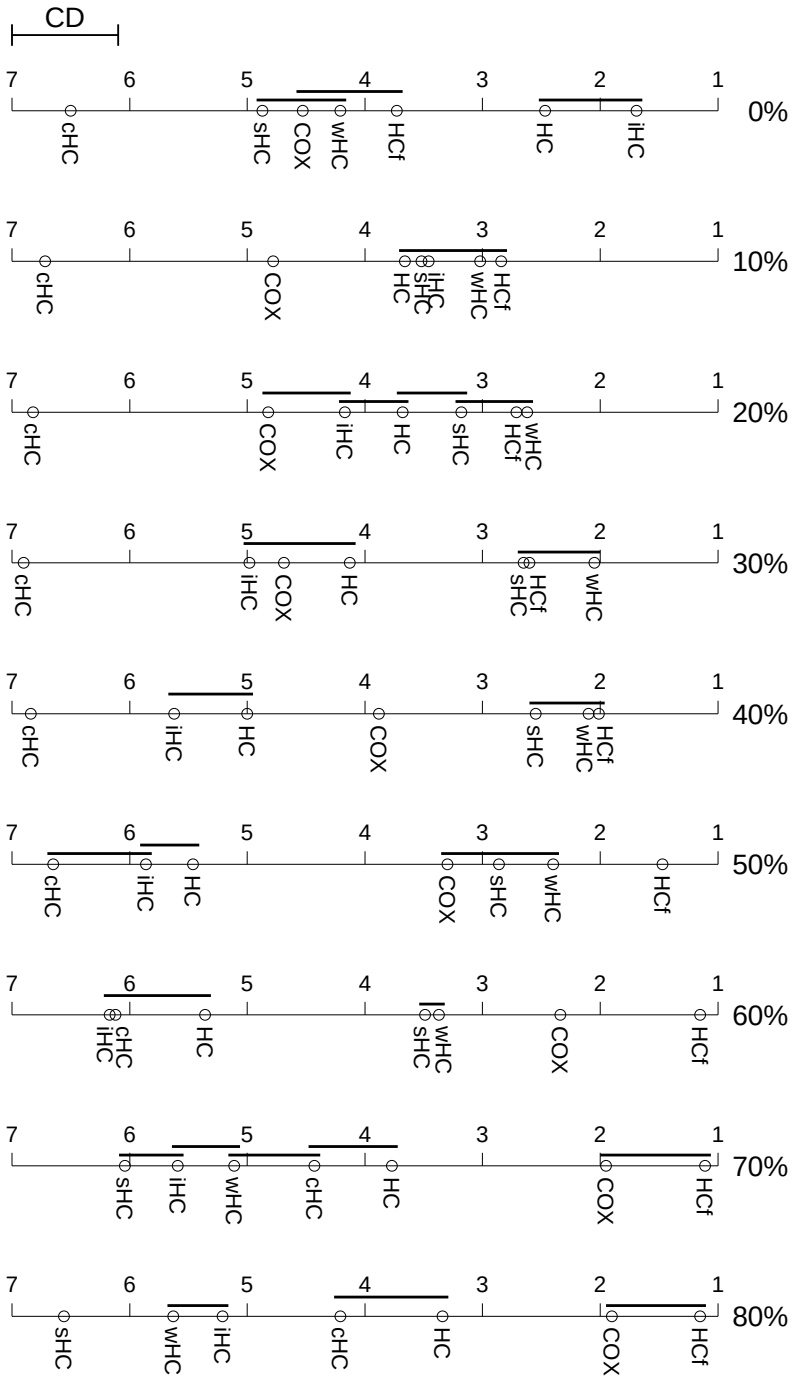


(b) Algoritam uvjetnih neovisnosti

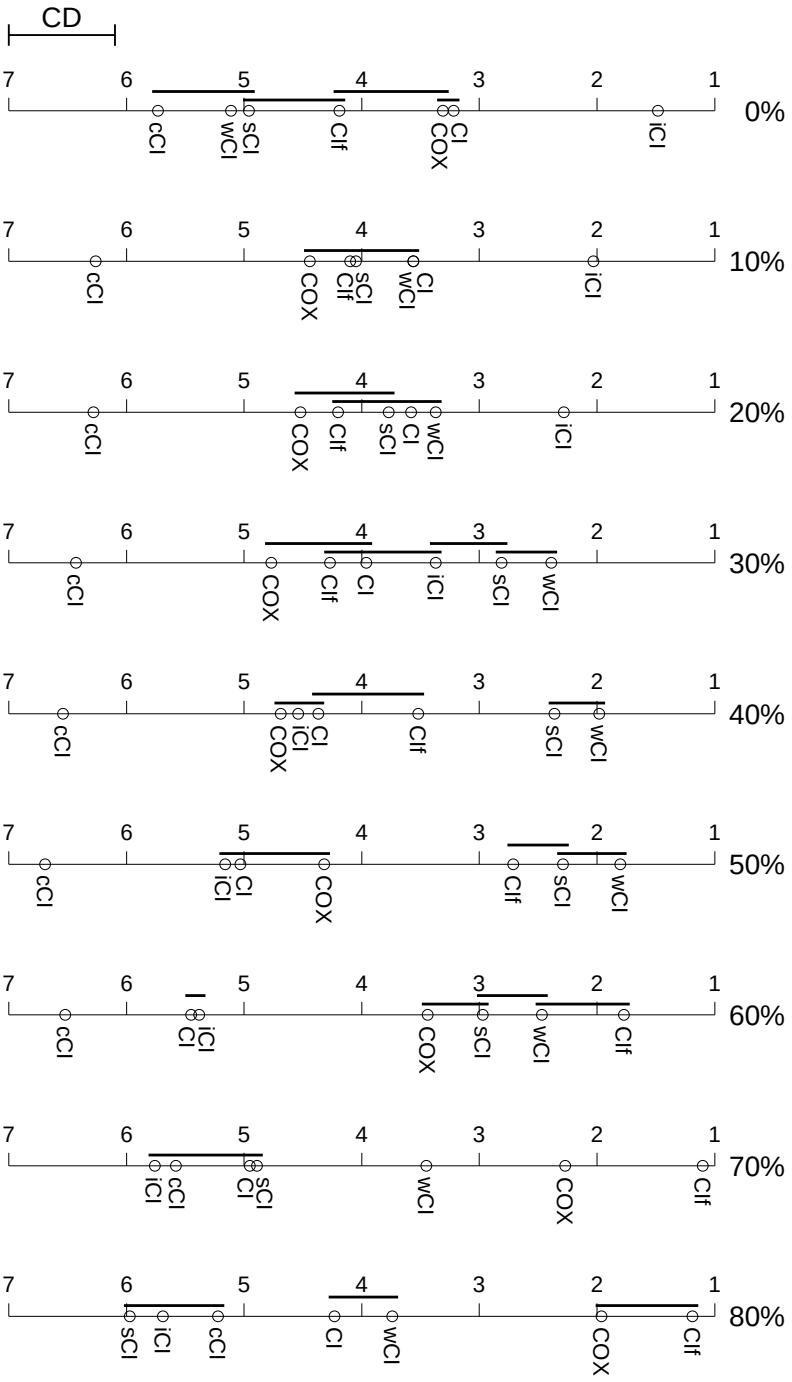


(c) Naivni Bayesov klasifikator

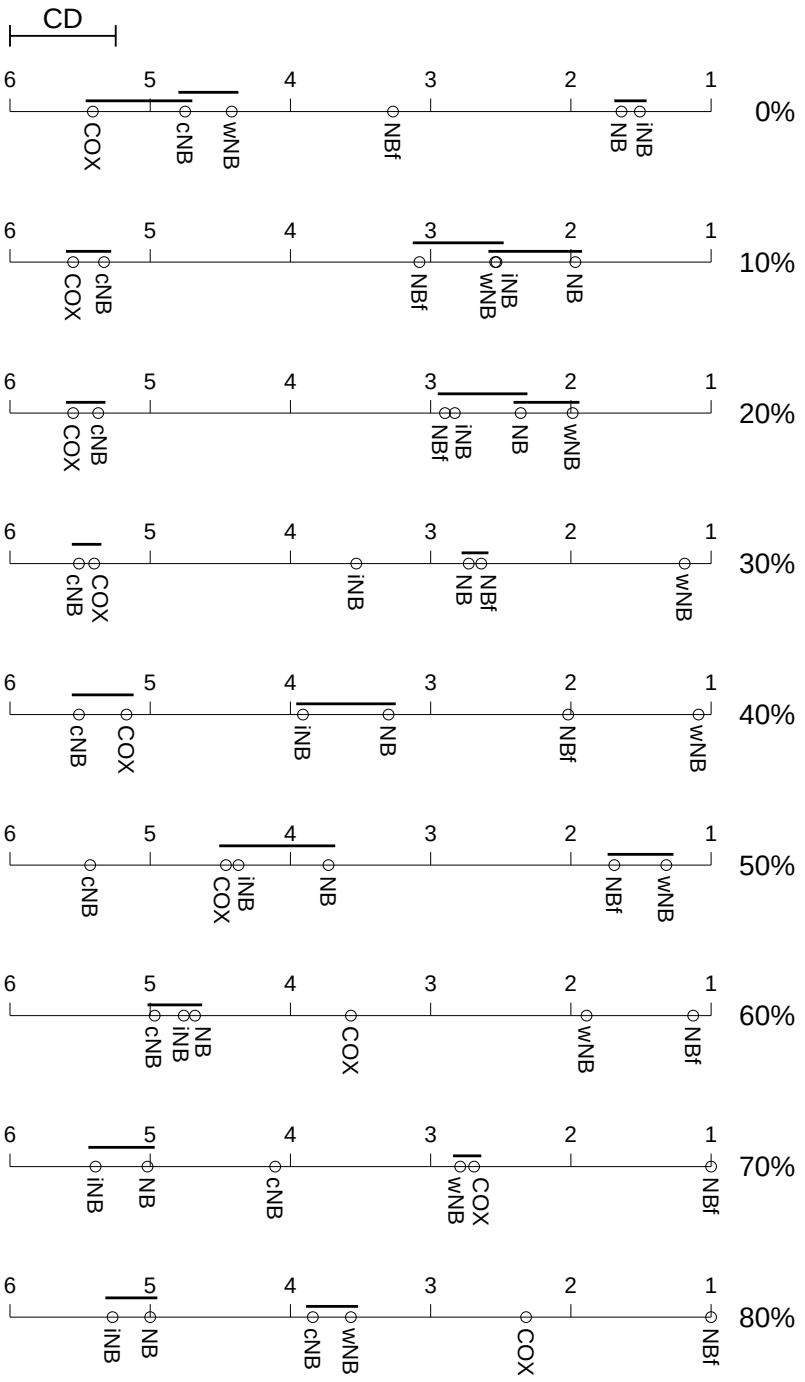
Slika 6.9: Rezidualna varijacija integrirane Brierove ocjene $\bar{x}$ sa standardnom devijacijom $\sigma(\bar{x})$.



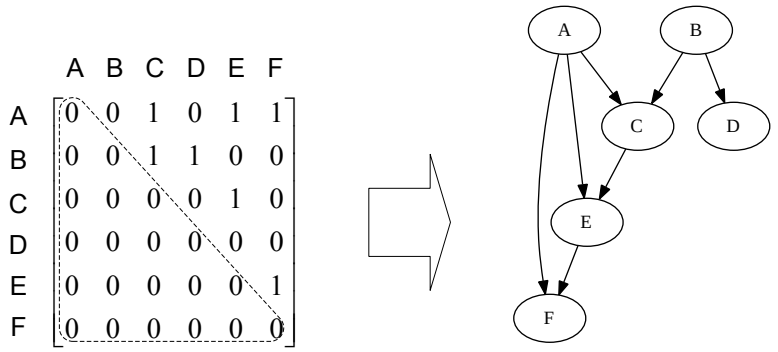
Slika 6.10: Srednji rangovi točnosti klasifikacije postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika 6.11: Srednji rangovi točnosti klasifikacije postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika 6.12: Srednji rangovi točnosti klasifikacije postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.

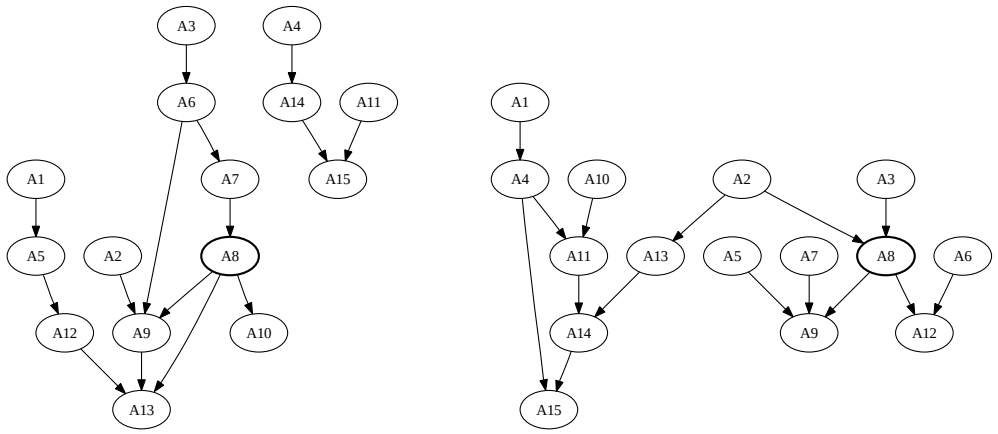


Slika 6.13: Primjer određivanja strukture grafa Bayesove mreže iz spojne matrice. Aciklič-nost grafa osigurana je ukoliko su svi elementi glavne dijagonale i svi elementi ispod glavne dijagonale spojne matrice jednaki nula.

jednu od kovarijata. U kojoj će mjeri kovarijate biti pogodne za predviđanje ishoda, utvrdi se strukturom mreže i tablicama uvjetnih distribucija vjerojatnosti. Upravo je ovaj broj čvorišta (kovarijata i ishoda) namjerno odabran kako bi generirani podaci predstavljali dovoljno složen problem vrijedan učenja, dok bi se inherentni problem istovremeno dao naučiti u realnom vremenu.

Lukovi među čvorištima mreže mogu se predstaviti spojnom matricom $\mathbf{C}$, to jest kvadratnom matricom veličine $N = 15$. Svaki element matrice $\mathbf{C}(i,j)$ ili je nula ili jedan, gdje ovo potonje predstavlja luk koji kreće iz čvorišta V_i te završava u čvorištu V_j (slika 6.13). Za svaki $V_i, i < o$ dodan je točno jedan luk; taj je usmjeren prema V_o s vjerojatnošću $\mathrm{P}(V_i \to V_o) = 0.3\overline{3}$ ili prema bilo kojem sljedećem čvorištu s vjerojatnošću $\mathrm{P}(V_i \to V_j) = 0.6\overline{6}, j > i$. Za svako čvorište $V_i, i > o$ dodan je točno jedan luk; taj kreće iz čvorišta V_o te je usmjeren prema čvorištu V_i s vjerojatnošću $\mathrm{P}(V_o \to V_i) = 0.3\overline{3}$ ili prema bilo kojem sljedećem čvorištu s vjerojatnošću $\mathrm{P}(V_i \to V_j) = 0.6\overline{6}, j > i$. Na koncu se dodaju još dva dodatna luka između bilo koja dva slučajno odabrana čvorišta $V_i \to V_j, i < j$, kako bi se povećala vjerojatnost da sva čvorišta budu (ne)posredno međusobno povezana. Ovakvim se postupkom naglašava povezivanje ostalih čvorišta s ishodnim čvorištem V_o, pritom zadovoljavajući kriterij da je generirana mreža usmjereni aciklični graf, to jest $\forall i \geq j : \mathbf{C}(i,j) = 0$. Izbor ishodnog čvorišta pao je na $o = 8$, kako bi se uspostavila ravnoteža između broja potencijalnih uzroka i posljedica ishoda. Kako bi se spriječilo generiranje suviše složenih mreža, to jest mreža koje nisu reprezentativne te kao takve nisu pogodne za učenje, postavljen je sljedeći uvjet - ukoliko je ijedno čvorište vezano s više od tri roditelja ili više od tri potomka, postupak generiranja mreže se ponavlja. Odabrani postupak generira modele koji naglašavaju interakciju kovarijata i ishoda, istovremeno zadržavajući jednostavnost strukture. Upravo je to svojstveno očekivanim prognostičkim modelima u kliničkoj medicini. Slika 6.14 prikazuje primjere generiranih mreža.

Po određivanju topologije mreže potrebno je odrediti oblik svakog kovarijatnog čvo-



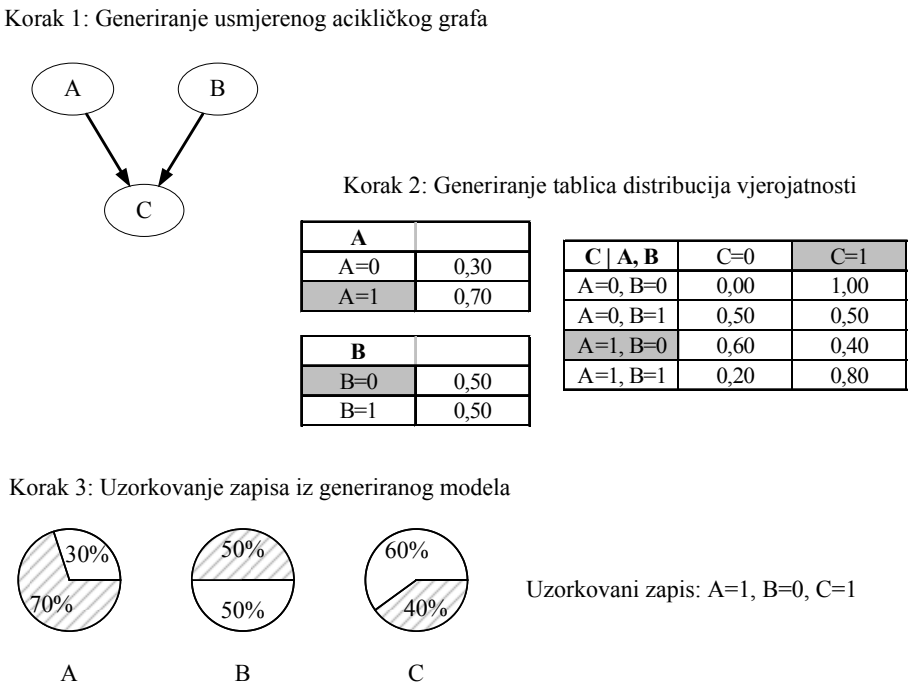
Slika 6.14: Topologije dviju slučajno generiranih Bayesovih mreža.

rišta te kvantificirati odnose povezanih čvorišta. Sve su kovarijate i ishod - diskretne binarne; moguće su vrijednosti "negativno" i "pozitivno". Sve su distribucije vjerojatnosti i uvjetne distribucije vjerojatnosti generirane slučajno korištenjem sljedećeg postupka. Za svako čvorište siroče (bez roditelja) V_i, generira se distribucija vjerojatnosti $(\mathrm{P}_i, 1-\mathrm{P}_i)$ uzorkovanjem vrijednosti p iz beta distribucije određene parametrima $\alpha = \beta = 0.2$. Za svako čvorište V_i s l roditelja, na isti se način generira ukupno 2^l međusobno neovisnih distribucija vjerojatnosti $(\mathrm{P}_i, 1-\mathrm{P}_i)$, po jedna za svaku moguću kombinaciju vrijednosti roditeljskih čvorišta.

6.2.2 Uzorkovanje i cenzuriranje primjeraka

Iz određene ishodišne Bayesove mreže uzorkuju se primjerci. Postupak uzorkovanja primjeraka započinje hijerarhijskim sortiranjem čvorišta mreže, čemu slijedi uzorkovanje pojedinačnih vrijednosti za svako čvorište, uzevši u obzir kombinaciju vrijednosti roditeljskih čvorišta (u slučaju da ona postoje). Postupak je ilustriran na slici 6.15. Kako bi se osigurao podjednak broj, prema ishodu pozitivnih i negativnih primjeraka u generiranom uzorku primjeraka, određeno je sljedeće ograničenje: ukoliko distribucija vjerojatnosti ishoda $(\mathrm{P}_o, 1-\mathrm{P}_o)$ prema generiranom uzorku ne zadovoljava kriterij $0.45 < \mathrm{P}_o < 0.55$, postupak generiranja ishodišnog modela i tablica uvjetnih distribucija vjerojatnosti se ponavlja.

Po okončanom postupku uzorkovanja podataka svakom je primjerku određeno vrijeme preživljenja i vrijeme promatranja te je na osnovi tih dviju vrijednosti određeno je li primjerak cenzuriran ili ne. Korišten postupak gotovo je istovjetan onom opisanom u sekciji 6.1.3 (Royston & Sauerbrei, 2004). Razlika se javlja u potrebi ovog postupka za izračunom regresijskih koeficijenata, koji tvore okosnicu izračuna vremena promatranja i vremena preživljenja, a koji su u prethodnoj simulacijskoj studiji bili određeni ishodišnim modelom (sekcija 6.1.1). Regresijski su koeficijenti $\boldsymbol{\beta}$ aproksimirani modelom logističke regresije (Hastie *et al.*, 2001), naučenom iz uzorkovanih podataka.



Slika 6.15: Ilustracija postupka uzorkovanja primjeraka nakon generiranja strukture mreže i tablica distribucija vjerojatnosti. Primjer je dan za mrežu od tri čvorišta. Za svaki se uzorkovani primjerak vrijednost svakog čvorišta odredi slučajnim odabirom, principom ruleta.

Postupak generiranja ishodišnog modela i uzorkovanja skupa primjeraka u potpunosti je ponovljen 100 puta; rezultati opisani u nastavku predstavljaju srednje vrijednosti i standardne devijacije za tih 100 uzorkovanih skupova podataka i njima pripadajućih modela. Svaki uzorak sadrži točno 1000 primjeraka. Eksperimentalno je utvrđeno da daljnjim povećanjem broja primjeraka rezultati vrednovanja nisu poboljšani.

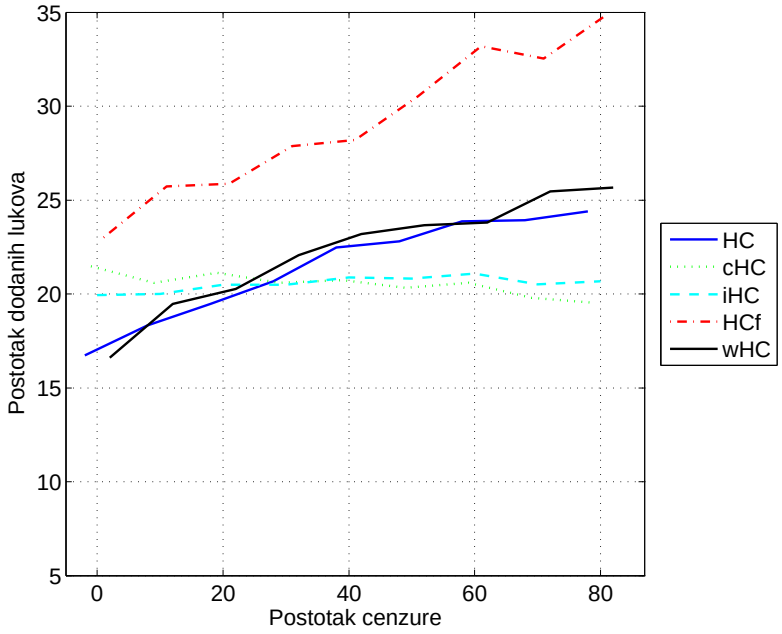
6.2.3 Rezultati

U nastavku su predstavljeni rezultati studije otkrivanja pravilnih topologija mreža. Pravilnost naučenih topologija, u odnosu na ishodišne, mjeri se brojem dodanih (slika 6.16), manjkajućih (slika 6.17) i obrnuto usmjerenih lukova (slika 6.18). Oznake korištenih postupaka opisane su u sekciji 5.3. Slike umjesto apsolutnog broja dodanih, manjkajućih ili obrnuto usmjerenih lukova prikazuju njihov postotak u odnosu na izvorni broj lukova (točno 15). Kod računanja broja obrnuto usmjerenih lukova uzimaju se u obzir samo oni lukovi koje je postupak uspješno detektirao, te ga je zato potrebno interpretirati u odnosu na broj nedostajućih lukova. Npr. malen broj obrnuto usmjerenih lukova ne znači ništa ako je broj manjkajućih lukova visok. Standardne devijacije navedenih mjera namjerno nisu predstavljene jer su svojom veličinom kvarile preglednost slika. Postupci podvajanja cenzuriranih primjeraka za učenje parametara mreže (sHC, sCI) namjerno su izostavljeni iz slika, jer su u učenju strukture jednaki postupcima koji tretiraju cenzurirane primjerke kao negativne (HC, CI). Za postupak učenja podjelom na vremenske intervale, simulacijom dobivene brojke predstavljaju aritmetičke sredine vrijednosti po naučenim modelima u svakom vremenskom intervalu.

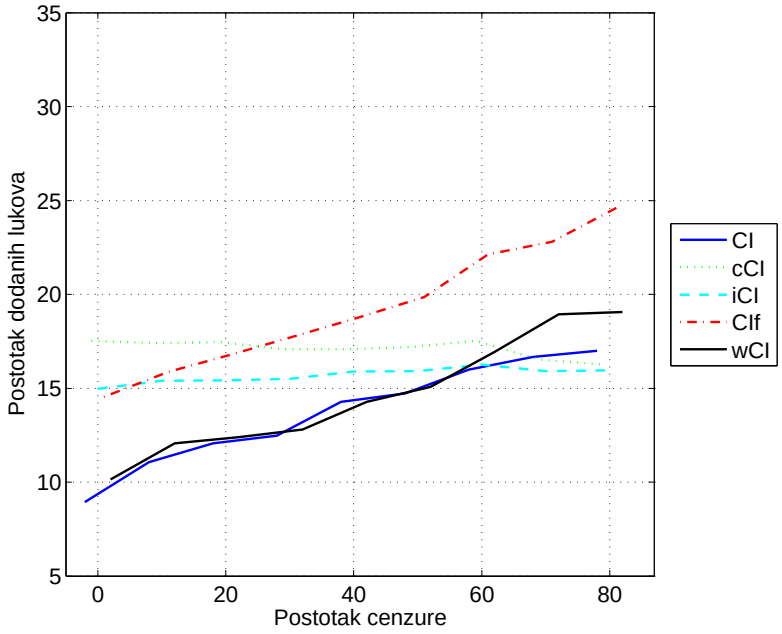
Za oba algoritma za učenje Bayesovih mreža postupak prilagodbe podataka odstranjivanjem šuma cenzure dodaje najveći broj suvišnih lukova; taj postupak, logično, istovremeno stvara modele s najmanje nedostajućih lukova. Modeli dobiveni postupkom učenja podjelom na vremenske intervale za oba algoritma pokazuju slične karakteristike modelima dobivenim postupkom prilagodbe podataka odstranjivanjem prekratko praćenih primjeraka (oba postupka rade s 500 ili manje primjeraka). Zanimljivo je da postupak podvajanja cenzuriranih primjeraka po performansama veoma sličan postupku tretiranja cenzuriranih primjeraka kao negativnih. Svi se postupci slično ponašaju s obrnutim usmjeravanjem lukova.

Navedene slike prikazuju postotke dodanih, manjkajućih ili obrnuto usmjerenih lukova u odnosu na cijelu ishodišnu mrežu. Detaljniji uvid u iste mjere, ali sada isključivo vezane uz razred (ishod), predstavljen je na slikama 6.19 (razredu dodani lukovi), 6.20 (razredu manjkajući lukovi) i 6.21 (razredu obrnuto usmjereni lukovi). Postupak prilagodbe podataka odstranjivanjem šuma cenzure razredu je dodao najviše lukova, ali ih istovremeno ima najmanje manjkajućih, što je konzistentno s rezultatima sa slika 6.16 i 6.17. Ostali se postupci po broju razredu dodanih lukova međusobno malo razlikuju. Kod visokog udjela cenzure postupak podjele na vremenske intervale (iHC, iCI) i postupak odstranjivanja prekratko praćenih primjeraka (cHC, cCI) razredu dodaju najmanje suvišnih lukova. Oba postupka, logično, imaju najviše razredu manjkajućih lukova na svim razinama cenzure.

Zbroj svih dodanih, manjkajućih i obrnuto usmjerenih lukova na nekom naučenom

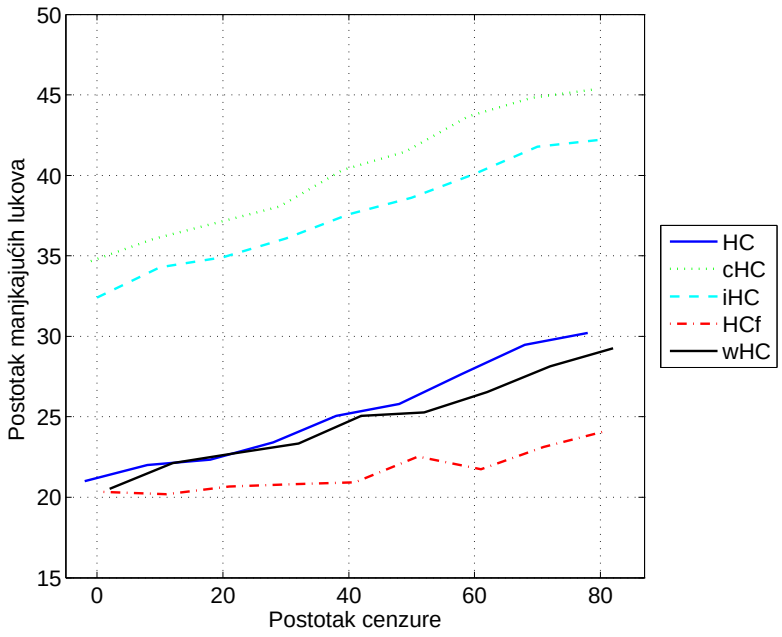


(a) Algoritam penjanja uzbrdo

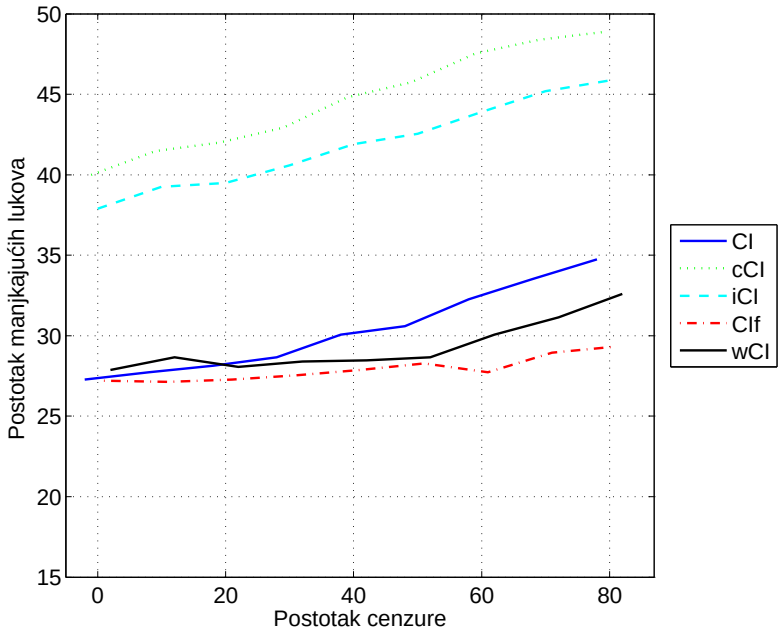


(b) Algoritam uvjetnih neovisnosti

Slika 6.16: Postotak dodanih lukova u odnosu na ishodišni model.

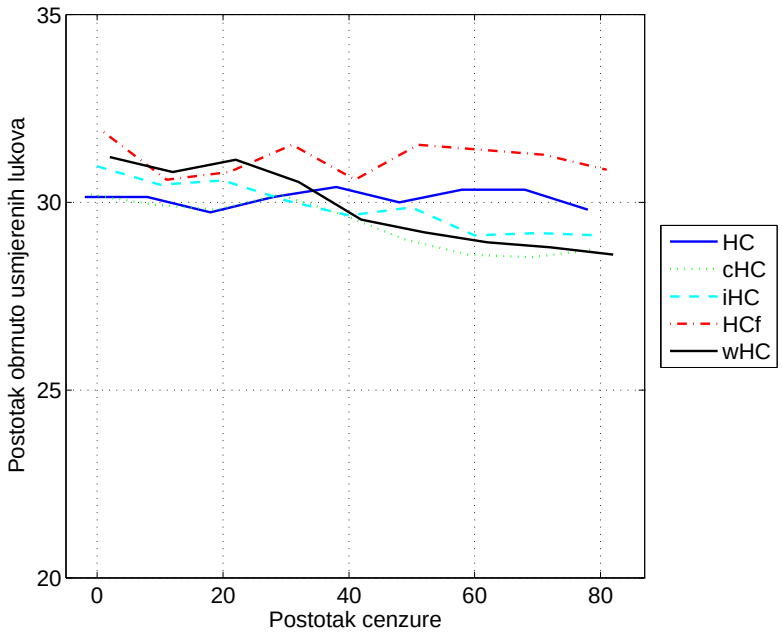


(a) Algoritam penjanja uzbrdo

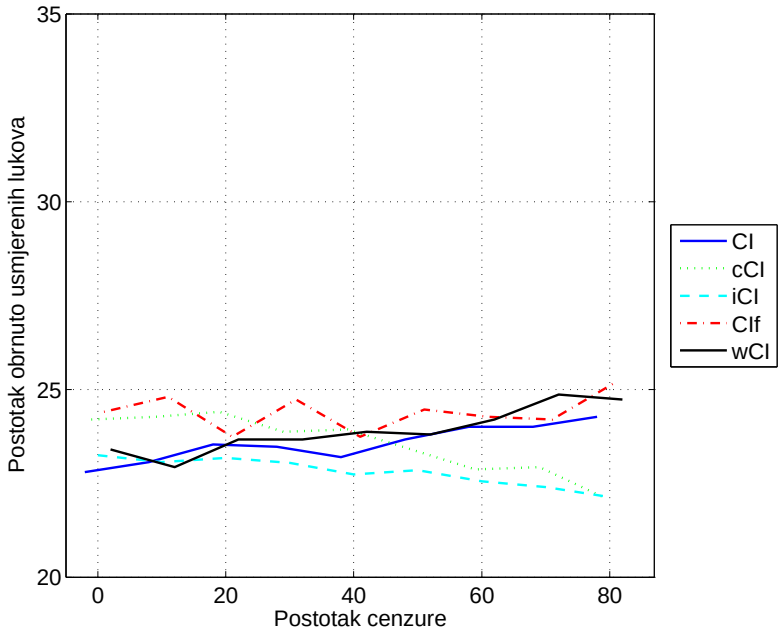


(b) Algoritam uvjetnih neovisnosti

Slika 6.17: Postotak manjkajućih lukova u odnosu na ishodišni model.

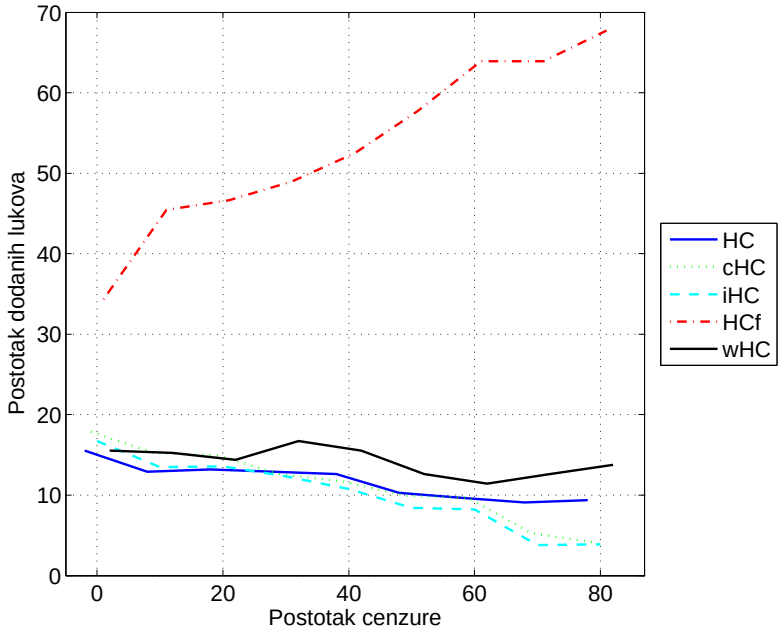


(a) Algoritam penjanja uzbrdo

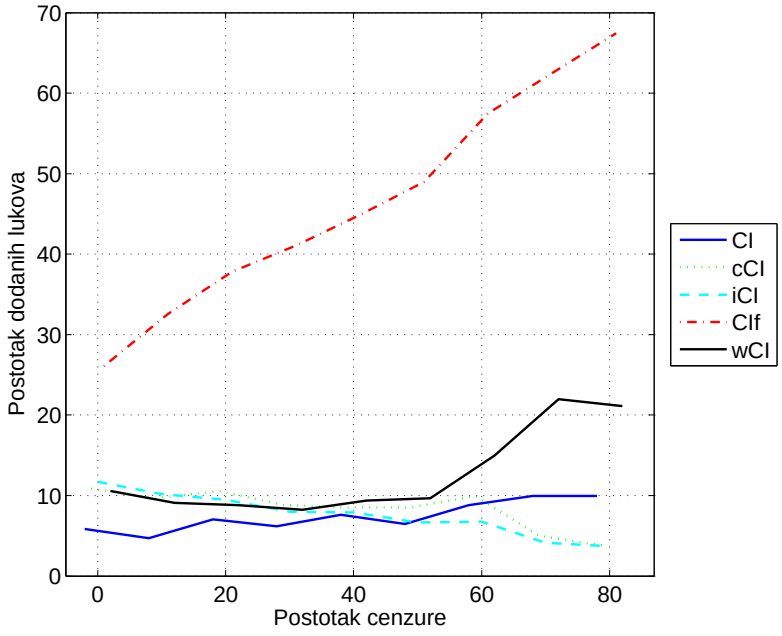


(b) Algoritam uvjetnih neovisnosti

Slika 6.18: Postotak obrnuto usmjerenih lukova u odnosu na ishodišni model.

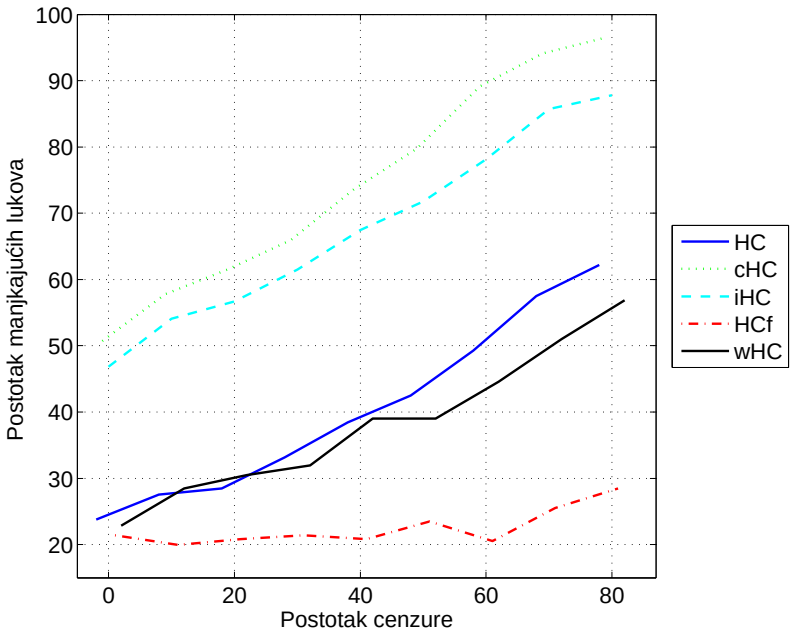


(a) Algoritam penjanja uzbrdo

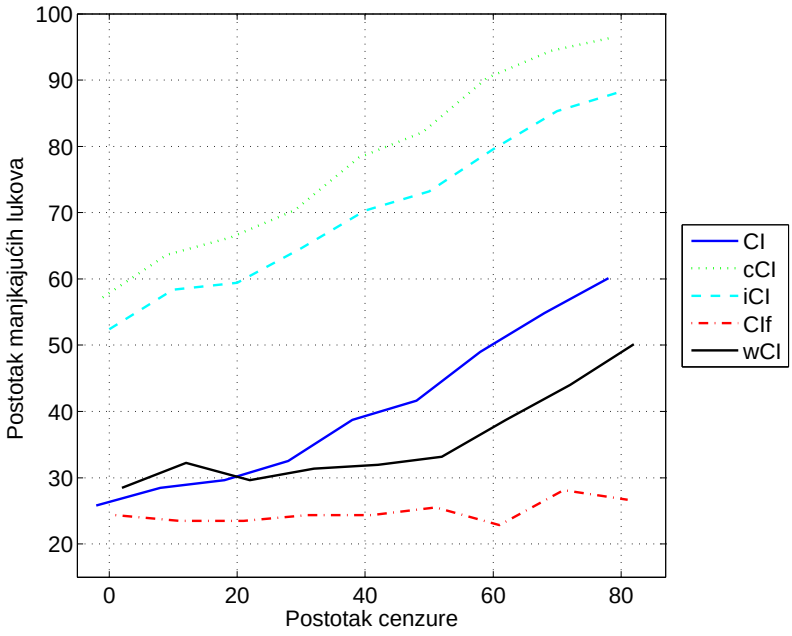


(b) Algoritam uvjetnih neovisnosti

Slika 6.19: Postotak razredu dodanih lukova u odnosu na ishodišni model.

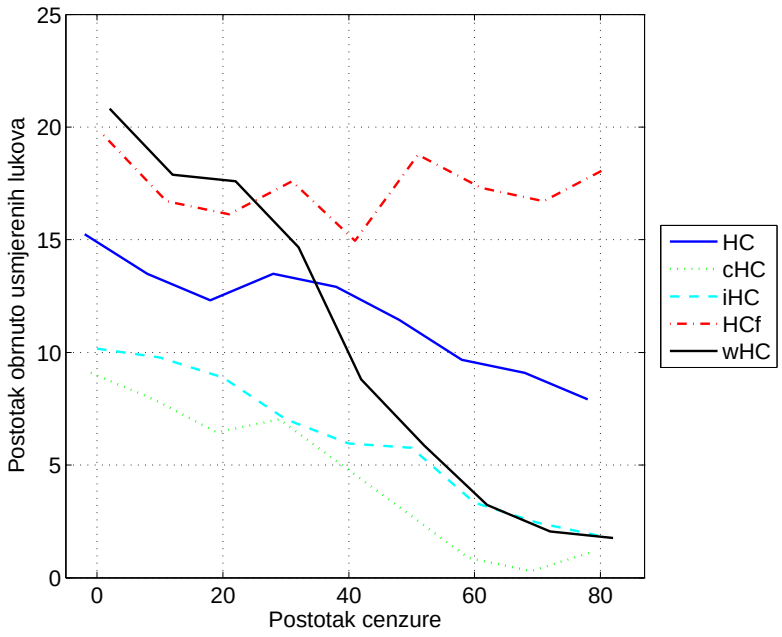


(a) Algoritam penjanja uzbrdo

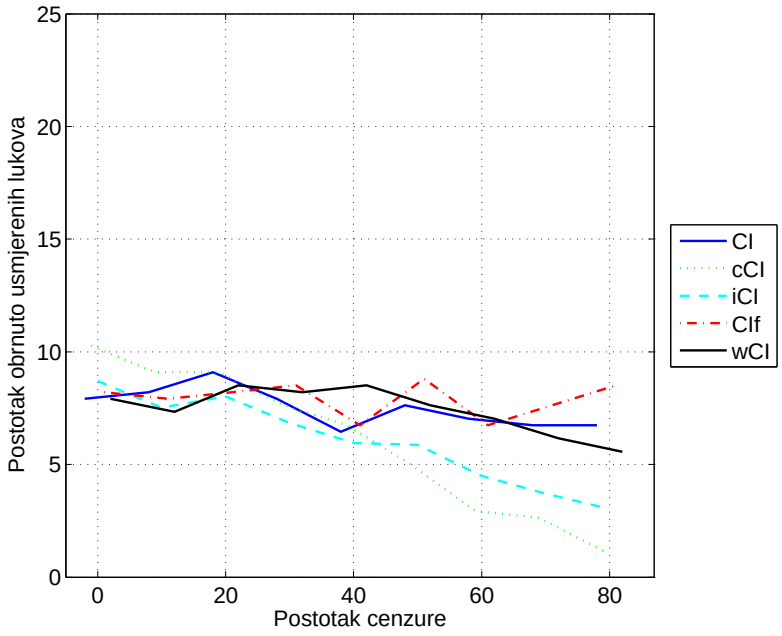


(b) Algoritam uvjetnih neovisnosti

Slika 6.20: Postotak razredu manjkajućih lukova u odnosu na ishodišni model.



(a) Algoritam penjanja uzbrdo

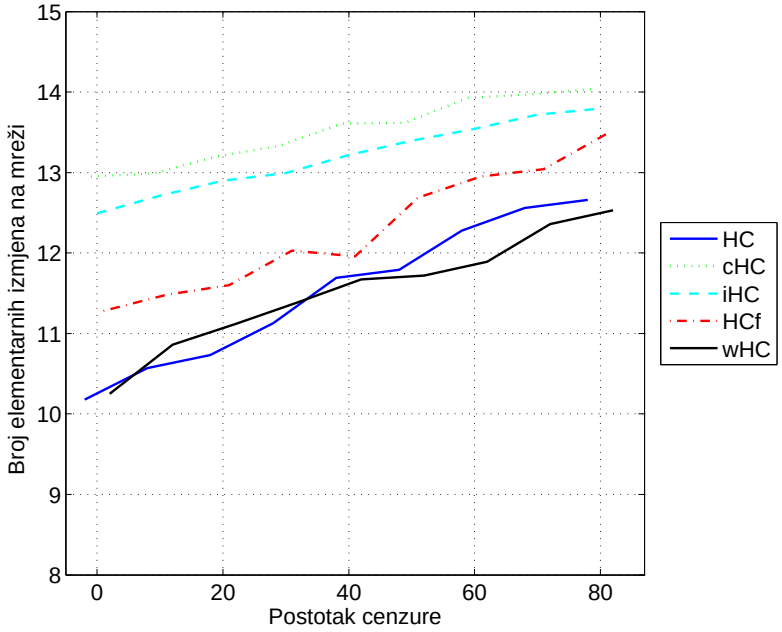


(b) Algoritam uvjetnih neovisnosti

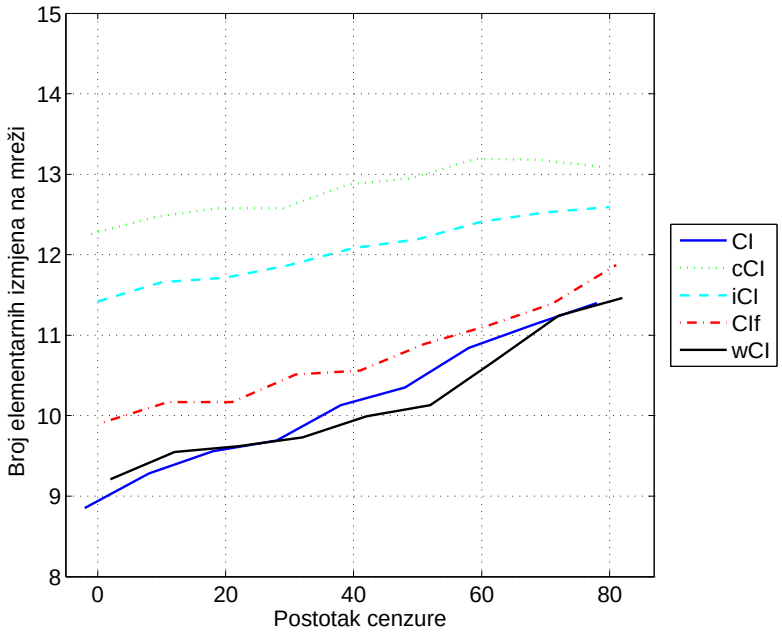
Slika 6.21: Postotak razredu obrnuto usmjerenih lukova u odnosu na ishodišni model.

modelu može se izraziti kao ukupan broj nužnih elementarnih operacija (izmjena mreže) za dostizanje topologije mreže ishodišnog modela. Ukupan broj elementarnih izmjena predstavljen je na slikama 6.22 (za cijelu mrežu) i 6.23 (za lukove neposredno spojene s razredom). Predstavljene krivulje ukazuju na to da modeli dobiveni postupkom učenja tretiranjem cenzuriranih primjeraka kao negativnih ili postupkom podvajanja cenzuriranih primjeraka uz težinske faktore, zahtijevaju najmanje elementarnih izmjena. Vidljivo je da je ovaj drugi u prosjeku nešto bolji za razine cenzure iznad približno 40%.

Srednji rangovi broja elementarnih izmjena na mreži postupaka učenja s različitim postupcima prilagodbe podataka za sve razine cenzure (od 0% do 80%) predstavljeni su na slikama 6.24 (algoritam penjanja uzbrdo) i 6.25 (algoritam uvjetnih neovisnosti). Srednji rangovi broja elementarnih izmjena na lukovima neposredno povezanim s razredom predstavljeni su na slikama 6.26 (algoritam penjanja uzbrdo) i 6.27 (algoritam uvjetnih neovisnosti). Za algoritam penjanja uzbrdo, postupci HC i wHC se ističu kao najbolji (statistički se razlikuju od ostalih na razini značajnosti testa $\alpha = 0.05$ za lukove vezane uz razred). Algoritam uvjetnih neovisnosti, pored postupaka CI i wCI, dodatno preferira i postupak CIf, ali statistički značajno ($\alpha = 0.05$) isključivo u elementarnim izmjenama na cijeloj mreži.

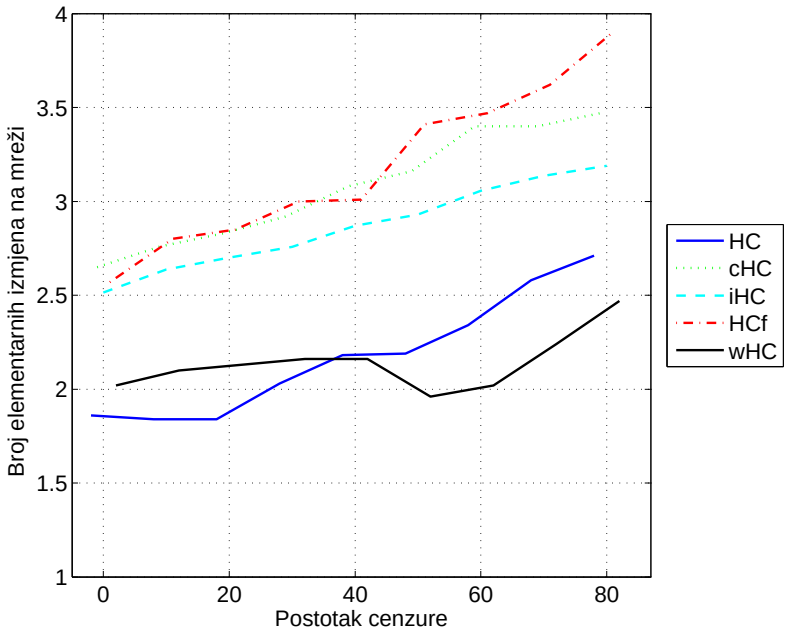


(a) Algoritam penjanja uzbrdo

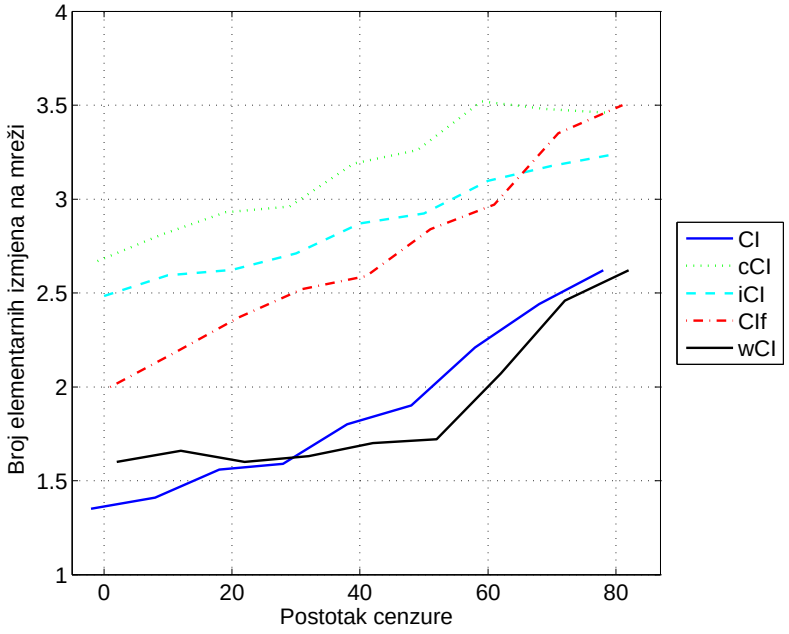


(b) Algoritam uvjetnih neovisnosti

Slika 6.22: Ukupan broj elementarnih izmjena (dodanih, manjkajućih ili obrnuto usmjerenih lukova) u odnosu na ishodišni model.

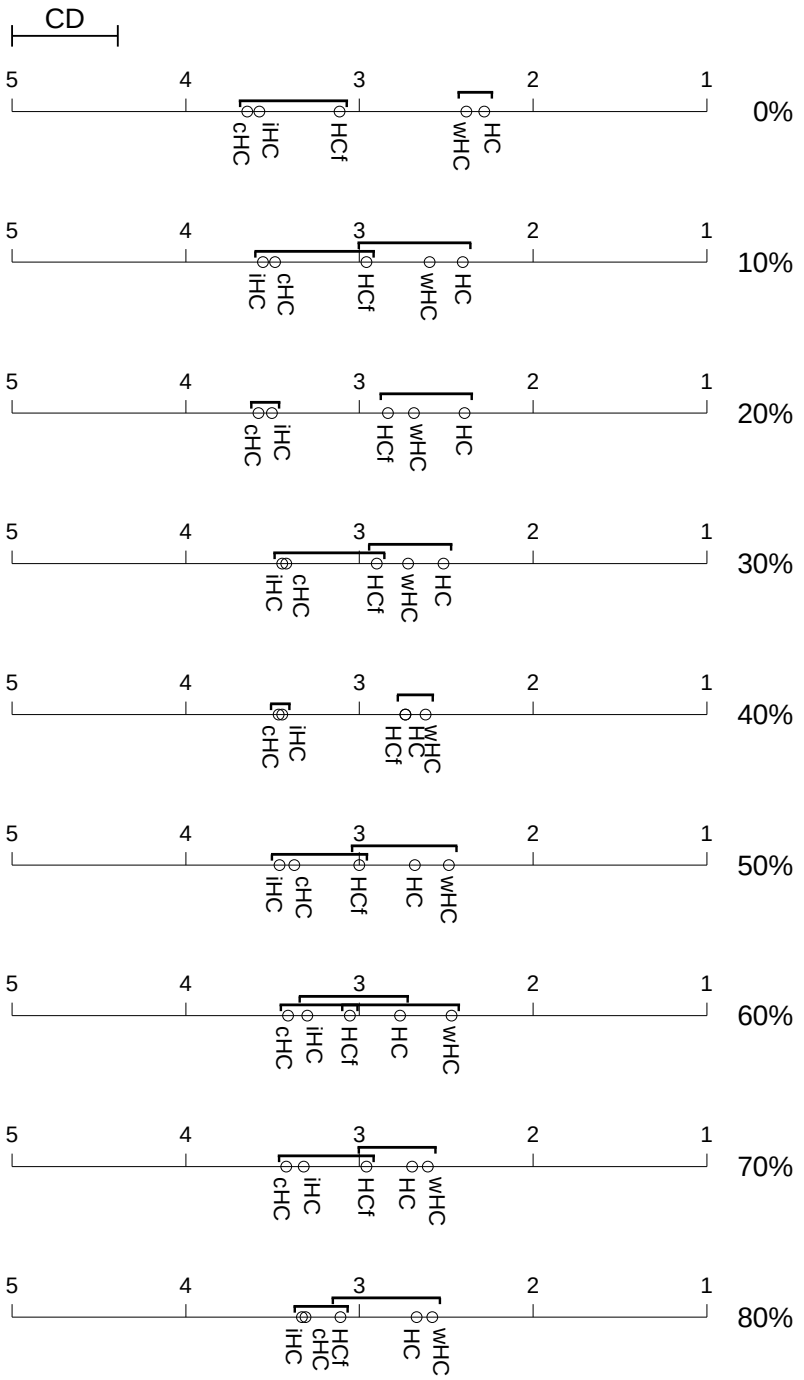


(a) Algoritam penjanja uzbrdo

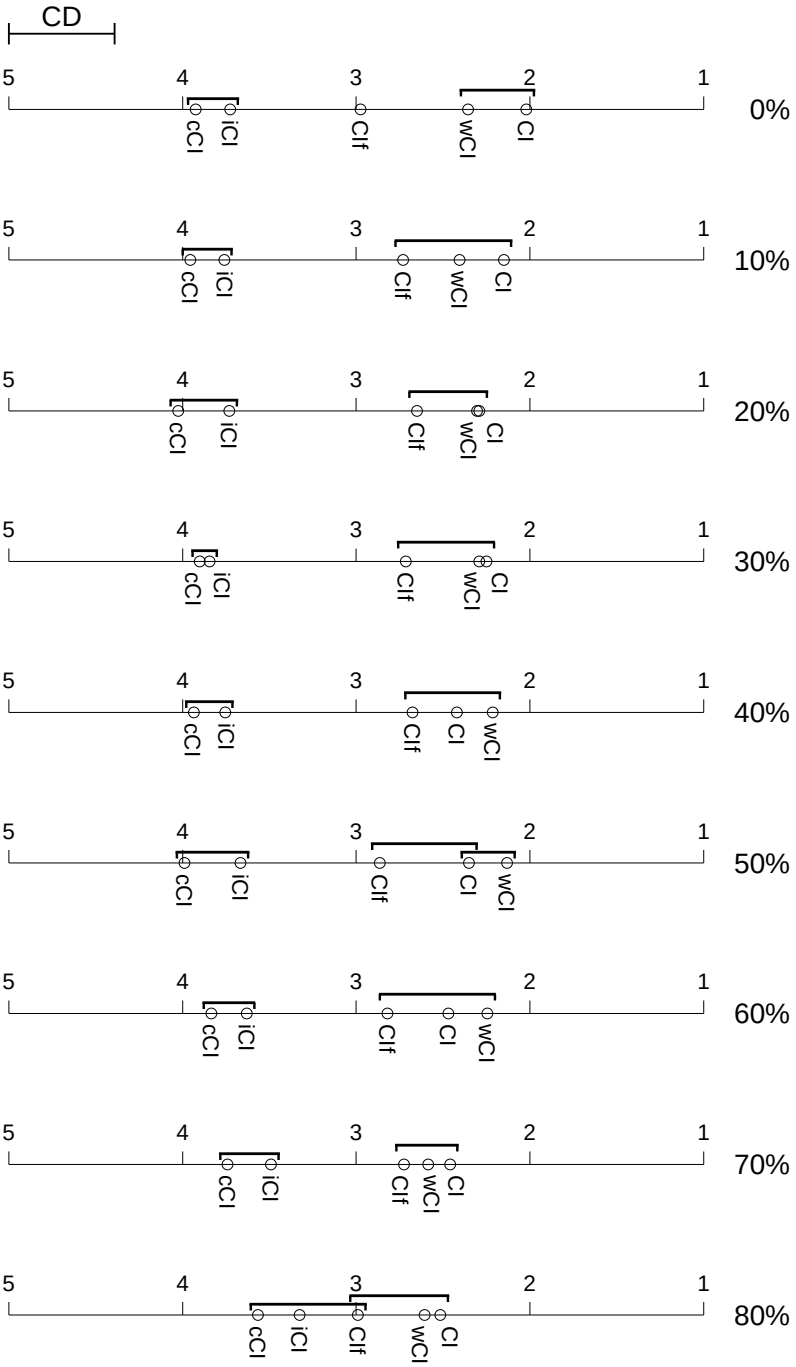


(b) Algoritam uvjetnih neovisnosti

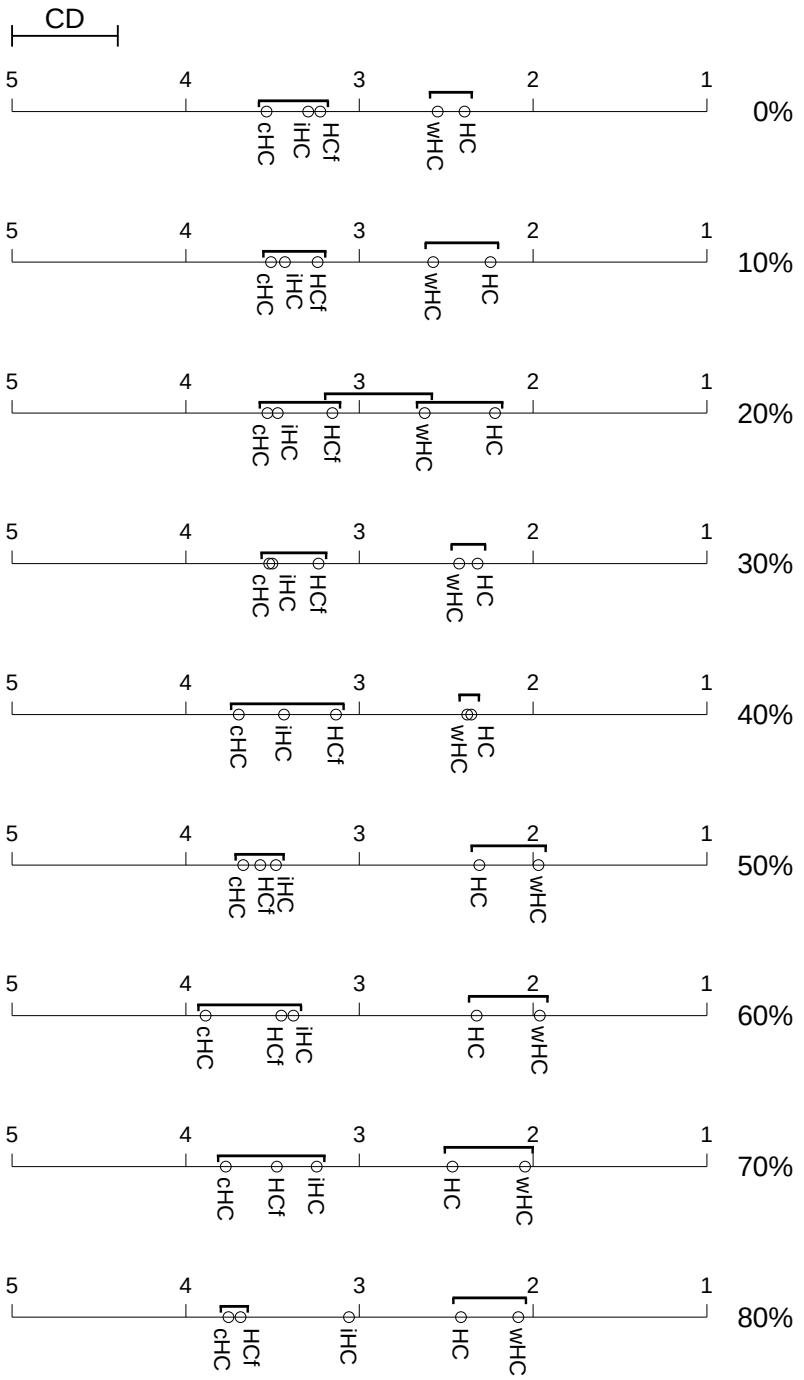
Slika 6.23: Ukupan broj elementarnih izmjena (dodanih, manjkajućih ili obrnuto usmjerenih lukova spojenih s razredom u ishodišnom modelu) u odnosu na ishodišni model.



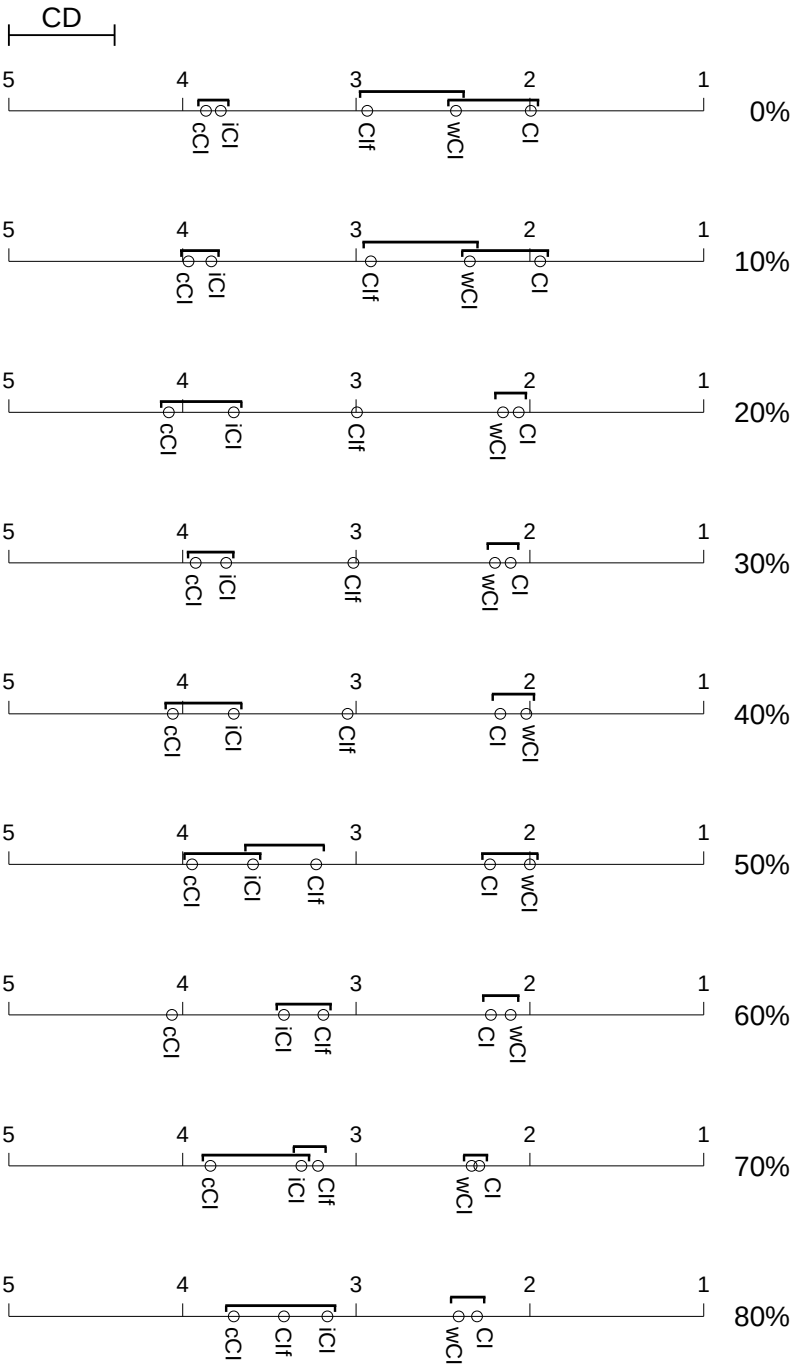
Slika 6.24: Srednji rangovi broja elementarnih izmjena na mreži postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika 6.25: Srednji rangovi broja elementarnih izmjena na mreži postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika 6.26: Srednji rangovi broja elementarnih izmjena uz razred na mreži postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika 6.27: Srednji rangovi broja elementarnih izmjena uz razred na mreži postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.

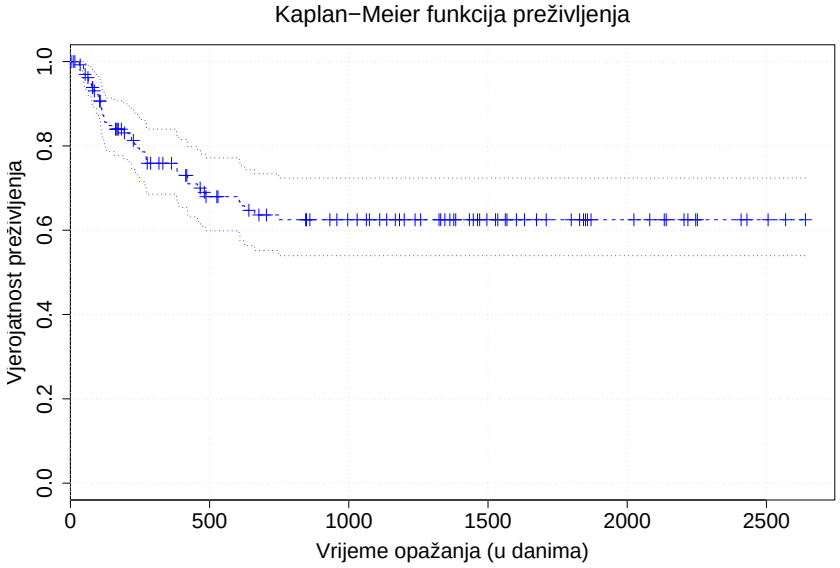
Poglavlje 7

Vrednovanje na realnim domenama

Utjecaj cenzure na postupke prilagodbe podataka za algoritme strojnog učenja i Coxovu regresiju, dodatno je analiziran na tri realne domene iz različitih područja kliničke medicine: transplantacije koštane srži (sekcija 7.1), ciroze jetre (sekcija 7.2) i raka dojke (sekcija 7.3). Područja su konceptualno različita po dimenzionalnosti i inherentnoj distribuciji, stoga su zanimljiva za dodatna testiranja opisanih postupaka. Navedeni podaci javno su dostupni i slobodni za upotrebu.

Oznake korištenih postupaka opisane su u sekciji 5.3. Za vrednovanje postupaka na domenama, primijenjene su sve opisane metrike. Vrijednosti predstavljene u tablicama srednje su vrijednosti deseterostruke stratificirane unakrsne provjere na deset iteracija. Dodatno su slikama predstavljene statističke sličnosti dobivenih vrijednosti relevantnih metrika uspoređenih postupaka. Za provjeru sličnosti postupaka korišten je Friedmanov dvostrani ANOVA test rangiranjem, kojemu su slijedili upareni neparametarski Wilcoxonovi dvostrani testovi rangiranih predznaka s Bonferronijevom korekcijom (sekcija 5.2.1). Rezultati nad kojima su se izvodili *post-hoc* testovi prikazani su Box i Whisker dijagramima sa sljedećom notacijom: crta u pravokutniku predstavlja medijanu, pravokutnik obuhvaća područje od prvog do trećeg kvartila, crte izvan pravokutnika na "brkovima" predstavljaju ekstremne rubne vrijednosti koje pripadaju uzorku, to jest one koje su od prvog ili trećeg kvartila udaljene najviše 150% interkvartilnog raspona, a križići predstavljaju vrijednosti koje ne pripadaju uzorku (engl. *outliers*). Grafički su sličnosti predstavljene tako da su postupci koji se ne razlikuju, povezani crtom (korekcija $\alpha = 0.05$). S obzirom na to da za navedene domene ne postoje gotovi ekspertni modeli u obliku Bayesovih mreža, nije bilo moguće provjeriti kakvoću strukture naučenih modela. Unatoč tome, u dodatku D su prikazane i Bayesove mreže naučene opisanim postupcima, kako bi čitatelj, bolje upoznat s navedenim područjima kliničke medicine, mogao dublje spoznati vrsnosti opisanih postupaka.

Svi su podaci prije učenja bili obrađeni na sljedeći način. Prvo, svi su nepotpuni primjerci (oni kojima neke značajke nisu poznate) bili uklonjeni. Drugo, sve su kontinuirane značajke bile diskretizirane. Zbog očekivanog cenzuriranja u podacima, standardni postupci nadzirane diskretizacije, poput postupka entropijske diskretizacije (Fayyad & Irani, 1993), nisu bili korišteni. Umjesto njih, korišten je postupak pred-



Slika 7.1: Funkcija preživljenja za domenu transplantacije koštane srži. Funkcija je praćena krivuljama 95%-tnih intervala pouzdanosti (crtkano).

ložen u Contal & O'Quigley (1999); Klein & Moeschberger (2003), koji se temelji na određivanju granice najboljeg log-rank razdvajanja. Taj postupak očito preferira naivni Bayesov klasifikator pred modelima Bayesovih mreža (uz algoritam penjanja uzbrdo i algoritam uvjetnih neovisnosti), jer preferira neposredne interakcije između svake kovarijate i ishoda. Iz tako diskretiziranih podataka, na jednoj od domena je algoritam za učenje regresijskog modela proporcionalnih hazarda dosegao maksimalan broj iteracija prije određivanja smislenog modela (konvergiranja procjene parametara). Iz tog su razloga posvuda predstavljeni i rezultati učenja regresijskog modela proporcionalnih hazarda iz izvorno nediskretiziranih podataka. Pritom je i tu korištena jednaka podjela na dijelove pri stratificiranoj unakrsnoj provjeri, kako bi rezultati bili usporedivi te kako bi statistički testovi bili ispravni.

7.1 Transplantacija koštane srži

Transplantacija koštane srži (engl. *bone marrow transplant*, BMT), poznata i kao postupak presađivanja matičnih stanica (engl. *stem cell transplant*), naziv je procesa uzimanja stanica iz koštane srži za reinfuziju u pacijenta poslije velikih doza kemoterapije ili radioterapije. Koštana srž je krvotvorno tkivo smješteno u moždinskim šupljinama kosti i u prostorima među gredicama spužvastog koštanog tkiva. U njoj nastaju sve krvne stanice, razgrađuju se eritrociti i stvara rezerva željeza nastala raspadanjem hemoglobina.

Transplantacija koštane srži jedan je od tretmana za liječenje akutne leukemije.

Od 1984. do 1989. godine, s najdužim praćenjem od 7 godina, u kliničkoj studiji je prikupljen uzorak od 137 pacijenata određenih za transplantaciju koštane srži (Klein & Moeschberger, 2003). Prilikom transplantacije, svaki je pacijent dodijeljen jednoj od tri rizične skupine: ALL (akutna limfoblastična leukemija), niskorizična AML (akutna mijelocitna leukemija) ili visokorizična AML. Svrha studije bila je praćenje preživljenja pacijenata bez bolesti, određene vremenom povrata, remisije ili smrti. Funkcija preživljenja predstavljena je na slici 7.1. Podaci su javno dostupni na adresi `http://www.mcw.edu/FileLibrary/Groups/Biostatistics/Publicfiles/DataFromSection/Data_from_section_1.3.doc` (dostupno: 21. prosinca 2009).

Pacijenti su opisani s 18 kovarijata, vremenom promatranja te pokazateljem povrata bolesti (status). Pokazatelj smrti bio je uklonjen iz podataka jer nije koristan za predviđanje povrata bolesti. Od preostalih 17 prognostičkih kovarijata diskretizirane su bile sljedeće: vremena do akutne i kronične reakcije odbacivanja organa (engl. *graft-vs-host disease*, GvHD), vrijeme do povrata broja trombocita na uobičajene vrijednosti, starost pacijenta i davaoca, te vrijeme čekanja na transplantaciju. Podaci su potpuni, to jest nema nedostajućih vrijednosti.

Odabrane granice za postupak prilagodbe podataka podjelom na vremenske intervale su $(0, 800, \infty)$. Rezultati stratificirane deseterostruke unakrsne provjere na deset iteracija predstavljeni su u tablici 7.1 (standardne metrike) i tablici 7.2 (metrike analize preživljenja). Postupak prilagodbe podataka podjelom na vremenske intervale po točnosti klasifikacije je na samom vrhu, za sva tri postupka učenja. Metrike analize preživljenja pokazuju da je za algoritam penjanja uzbrdo, najbolji predloženi postupak prilagodbe podataka odstranjivanjem šuma cenzure, dok su za algoritam uvjetnih neovisnosti i naivni Bayesov klasifikator, podjednako uspješni podjela na vremenske intervale i podvajanje cenzuriranih primjeraka.

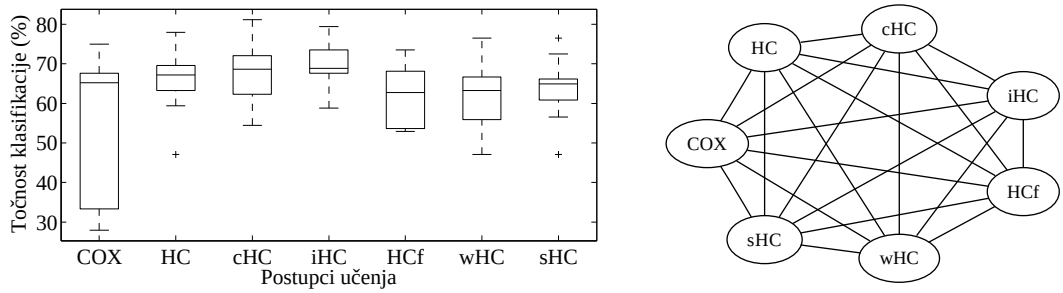
Statistički testovi predstavljeni su na slikama 7.2 (algoritam penjanja uzbrdo), 7.3 (algoritam uvjetnih neovisnosti) i 7.4 (naivni Bayesov klasifikator). Postupak odstranjivanja prekratko praćenih primjeraka u većini je slučajeva statistički lošiji od postupka podjele na vremenske intervale (za sva tri modela) i od postupka podvajanja cenzuriranih primjeraka (osim za HC). Ni jedan se drugi par postupaka sustavno (po većini metrika) statistički značajno ne razlikuje.

Tablica 7.1: Vrednovanje različitih postupaka učenja iz domene transplantacije koštane srži. Korištene metrike su točnost klasifikacije, osjetljivost i specifičnost. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

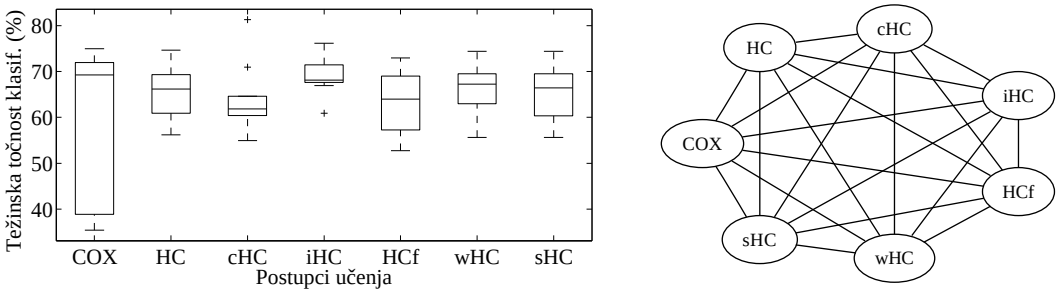
	Mjera					
Postupak	Točnost klasifikacije (%)		Osjetljivost (%)		Specifičnost (%)	
Penjanje uzbrdo						
HC	69.42	(3.67)	39.52	(6.75)	82.63	(3.85)
cHC	69.34	(2.41)	24.05	(10.31)	**89.37**	(4.75)
iHC	**71.46**	(3.53)	39.52	(6.75)	85.58	(3.40)
HCf	66.28	(1.57)	**74.76**	(3.40)	62.53	(2.12)
wHC	65.69	(1.88)	59.29	(9.81)	68.53	(5.62)
sHC	69.34	(2.20)	58.81	(8.48)	74.00	(5.11)
Uvjetne neovisnosti						
CI	71.02	(2.41)	53.81	(4.65)	78.63	(3.02)
cCI	68.47	(2.77)	20.00	(7.38)	**89.89**	(2.64)
iCI	**72.70**	(2.34)	53.81	(4.65)	81.05	(2.98)
CIf	63.21	(3.14)	**75.71**	(3.33)	57.68	(4.03)
wCI	65.69	(2.23)	69.05	(2.51)	64.21	(3.02)
sCI	69.05	(1.62)	66.67	(3.37)	70.11	(2.28)
Naivni Bayesov klasifikator						
NB	71.82	(0.92)	49.05	(2.30)	81.89	(1.09)
cNB	67.81	(1.67)	23.10	(3.18)	**87.58**	(1.55)
iNB	**71.90**	(0.86)	49.05	(2.30)	82.00	(1.05)
NBf	67.66	(1.98)	**73.81**	(5.26)	64.95	(1.65)
wNB	69.05	(1.51)	64.29	(3.72)	71.16	(1.02)
Proporcionalni hazard						
COX	**68.54**	(2.29)	**65.48**	(2.57)	69.89	(4.04)
COXc	64.96	(1.65)	52.62	(4.27)	**70.42**	(2.35)

Tablica 7.2: Vrednovanje različitih postupaka učenja iz domene transplantacije koštane srži. Korištene metrike su indeks suglasnosti, težinska točnost klasifikacije, integrirana Brierova ocjena i njena rezidualna varijacija. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

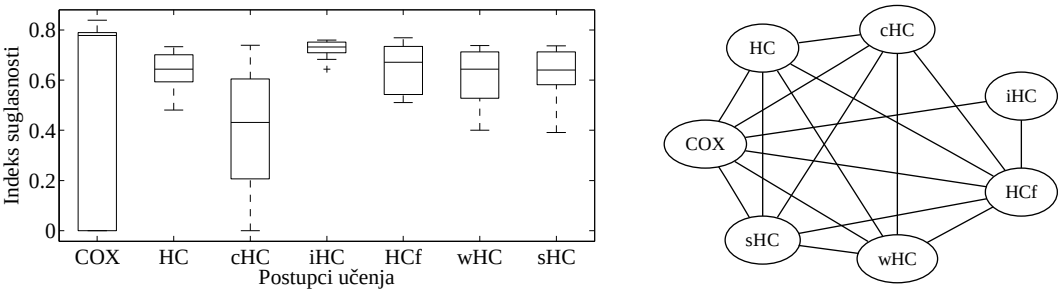
	Mjera							
Postupak	Težinska toč. klasif. (%)		Indeks suglasnosti		Integrirana Brierova ocjena		Rezidualna varijacija	
Penjanje uzbrdo								
HC	66.7	(3.0)	0.660	(0.042)	0.118	(0.005)	14.1	(3.1)
cHC	65.9	(2.4)	0.534	(0.059)	0.132	(0.010)	3.8	(6.9)
iHC	68.8	(2.9)	0.743	(0.041)	0.114	(0.005)	17.2	(3.0)
HCf	**69.9**	(1.2)	**0.751**	(0.041)	**0.113**	(0.004)	**18.0**	(3.3)
wHC	67.7	(1.7)	0.665	(0.044)	0.117	(0.006)	15.1	(2.8)
sHC	69.1	(1.8)	0.663	(0.046)	0.115	(0.004)	16.1	(2.9)
Uvjetne neovisnosti								
CI	69.9	(1.7)	0.730	(0.026)	0.117	(0.005)	15.2	(2.0)
cCI	64.8	(2.2)	0.542	(0.046)	0.138	(0.011)	-0.6	(5.8)
iCI	**71.6**	(1.6)	0.764	(0.028)	0.113	(0.005)	17.8	(2.1)
CIf	67.8	(2.7)	0.747	(0.025)	0.119	(0.004)	13.4	(2.9)
wCI	69.6	(1.5)	**0.766**	(0.021)	**0.109**	(0.002)	**20.4**	(2.3)
sCI	70.6	(1.1)	0.731	(0.032)	0.115	(0.005)	16.6	(2.0)
Naivni Bayesov klasifikator								
NB	71.3	(0.7)	0.806	(0.022)	0.102	(0.003)	26.1	(1.7)
cNB	65.9	(1.3)	0.772	(0.024)	0.119	(0.006)	13.3	(3.4)
iNB	71.4	(0.7)	**0.829**	(0.020)	0.100	(0.003)	27.3	(1.7)
NBf	71.6	(1.9)	0.803	(0.025)	0.103	(0.004)	25.2	(2.8)
wNB	**72.5**	(1.3)	0.820	(0.027)	**0.095**	(0.003)	**30.9**	(1.7)
Proporcionalni hazard								
COX	**72.4**	(1.8)	**0.790**	(0.039)	**0.099**	(0.008)	**27.9**	(6.7)
COXc	68.2	(1.5)	0.771	(0.046)	0.104	(0.009)	25.1	(6.7)



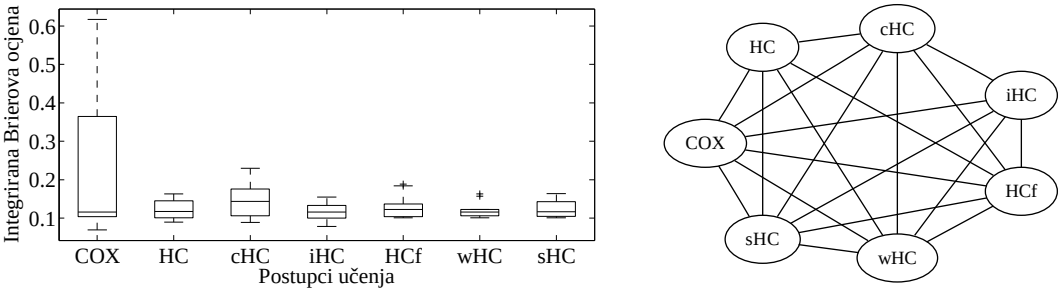
(a) Točnost klasifikacije



(b) Težinska točnost klasifikacije

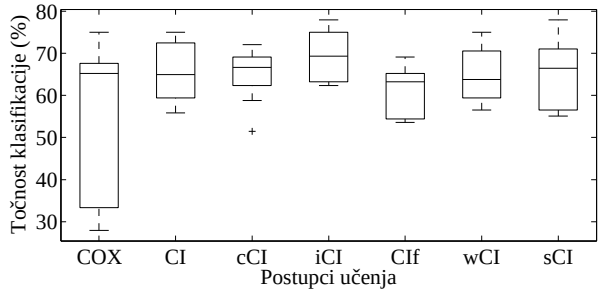


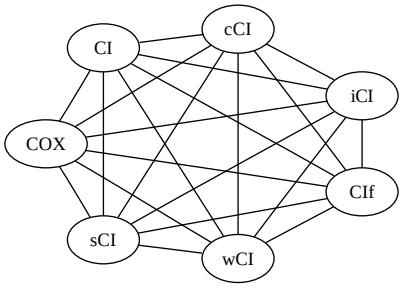
(c) Indeks suglasnosti



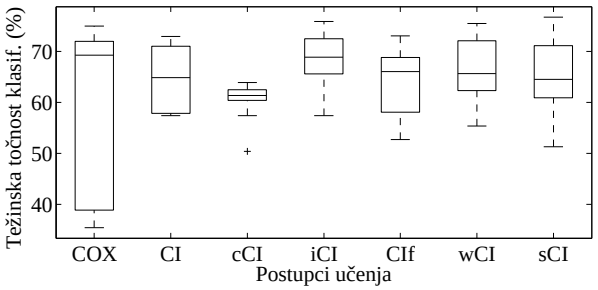
(d) Integrirana Brierova ocjena

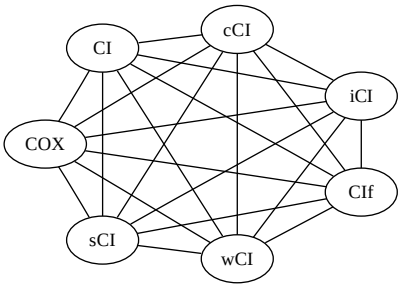
Slika 7.2: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena transplantacije koštane srži).



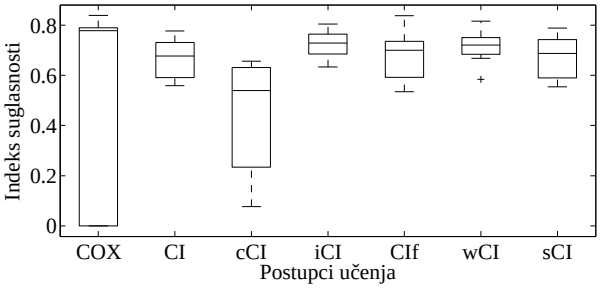


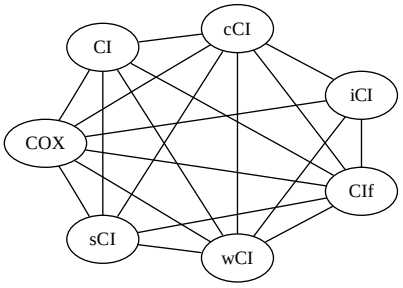
(a) Točnost klasifikacije



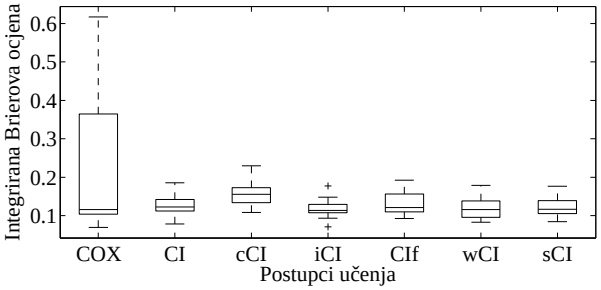


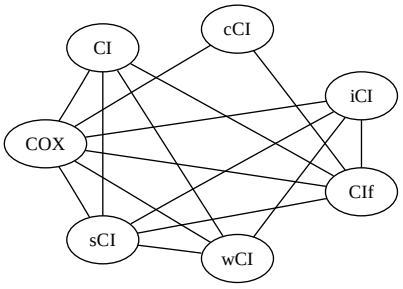
(b) Težinska točnost klasifikacije





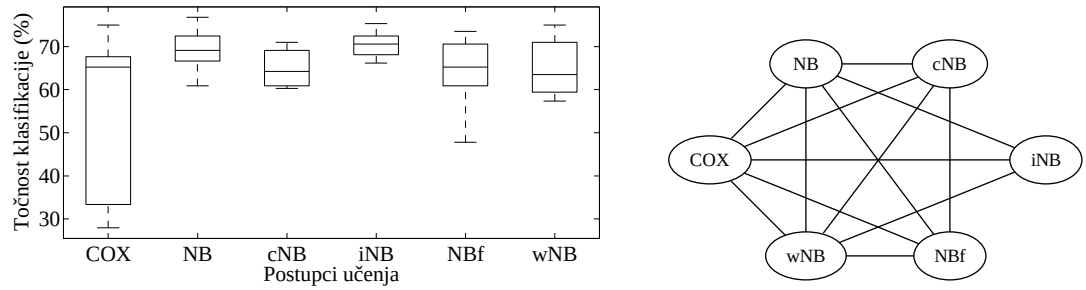
(c) Indeks suglasnosti



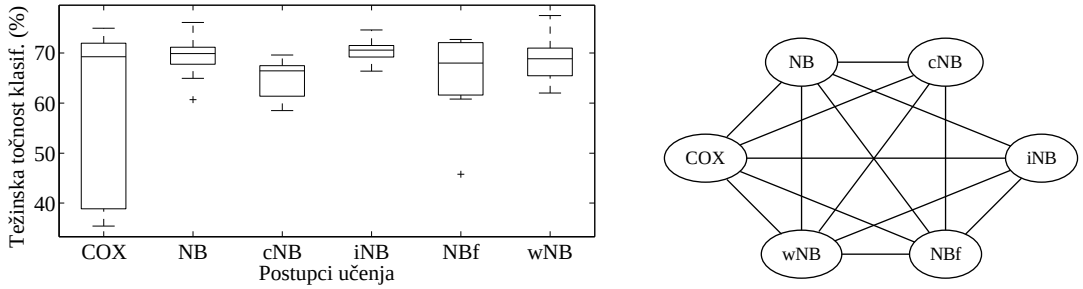


(d) Integrirana Brierova ocjena

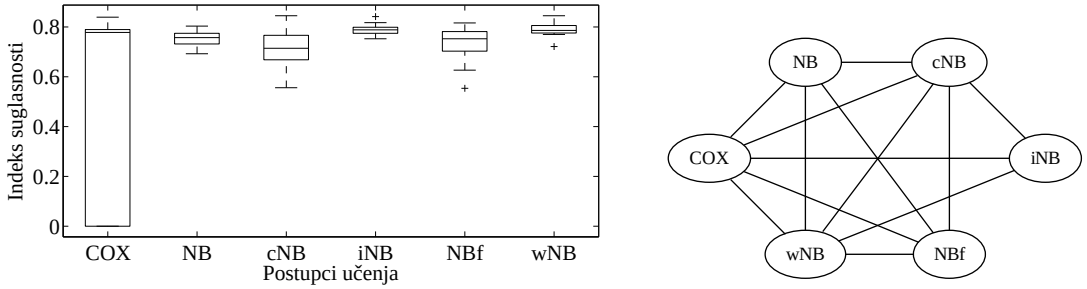
Slika 7.3: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena transplantacije koštane srži).



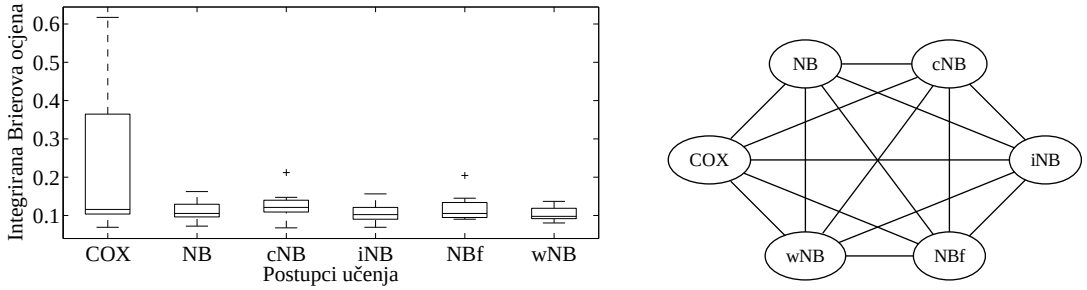
(a) Točnost klasifikacije



(b) Težinska točnost klasifikacije

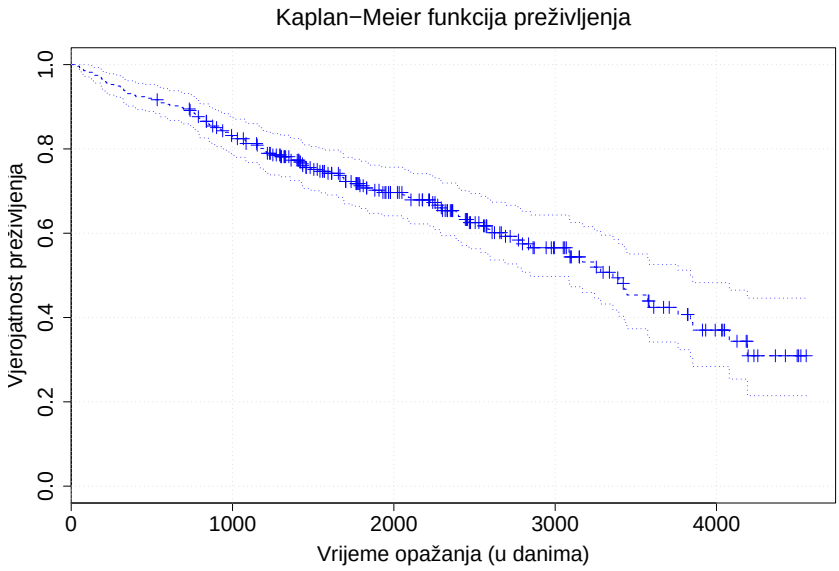


(c) Indeks suglasnosti



(d) Integrirana Brierova ocjena

Slika 7.4: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena transplantacije koštane srži).



Slika 7.5: Funkcija preživljenja za domenu ciroze jetre. Funkcija je praćena krivuljama 95%-tnih intervala pouzdanosti (crtkano).

7.2 Ciroza jetre

Ciroza jetre je kronična bolest jetre tijekom koje se jetreno tkivo zamjenjuje vezivnim tkivom što za krajnju posljedicu ima prestanak funkcija jetre. Ciroza jetre može imati cijeli niz uzroka od kojih su najčešći alkoholizam i hepatitis C. S obzirom da se oštećeno jetreno tkivo ne može zamijeniti, terapija ciroze jetre je palijativne prirode, iako se u ekstremnim situacijama može provesti transplantacija jetre. Primarna bilijarna ciroza je kronična i progresivna bolest jetre, nepoznate etiologije (vjerojatno autoimune), od koje najčešće obolijevaju žene srednje životne dobi.

Klinička studija primarne bilijarne ciroze jetre (engl. *primary billiary cirrhosis*, PBC) provedena je na klinici Mayo između 1974. i 1984. godine, uz praćenje do 1988. godine. Ukupno su 424 PBC-pacijenta zadovoljila kriterije randomizirane placebo-kontrolirane studije lijeka D-penicilamina. Od tog broja pacijenata, njih 312 je pristalo sudjelovati u studiji, uz potvrdu liječnika. Za vrijeme trajanja studije i praćenja, 125 od 312 pacijenata je umrlo. Od toga broja smrti, njih 11 nije se moglo pripisati PBC-u. Pored toga, 8 pacijenata je bilo izgubljeno za vrijeme praćenja, dok ih je 19 bilo podvrgnuto transplantaciji jetre (Fleming & Harrington, 1991). Klinička je studija osporila utjecaj lijeka D-penicilamina na ishod liječenja pacijenata, stoga su se podaci iskoristili za proučavanje prirodne povijesti bolesti. Funkcija preživljenja predstavljena je na slici 7.5. Podaci su javno dostupni na adresi `http://lib.stat.cmu.edu/S/Harrell/data/descriptions/pbc.html` (dostupno: 21. prosinca 2009).

Svaki zapis o pacijentu sastoji se od 16 prognostičkih kovarijata, tretmana, vremena promatranja i statusa. Status je pokazivačka varijata koja opisuje je li promatranje

pacijenta bilo cenzurirano ili je pacijent zaista umro iz razloga vezanih uz PBC. Zapisi 36 pacijenata bili su isključeni zbog nedostajućih vrijednosti. Konačni skup podataka sastojao se iz 111 pacijenata koji su umrli iz razloga vezanih uz PBC i 165 pacijenata koji su bili cenzurirani. Cilj je bio istražiti utjecaj prognostičkih kovarijata i lijeka D-penicilamina na ishod bolesti.

Sljedeće kovarijate su bile diskretizirane: starost, bilirubin, kolesterol, albumin, sediment urina, alkalna fosfataza, serumska glutamat-oksaloacetat transaminaza (SGOT), trigliceridi, razina trombocita i protrombina.

Odabrane granice za postupak prilagodbe podataka podjelom na vremenske intervale su $(0, 1000, 2000, 3000, \infty)$. Rezultati stratificirane deseterostruke unakrsne provjere na deset iteracija predstavljeni su u tablici 7.3 (standardne metrike) i tablici 7.4 (metrike analize preživljenja). Za oba algoritma učenja Bayesovih mreža po točnosti klasifikacije najbolji je postupak podjele na vremenske intervale. Od metrika analize preživljenja jedino indeks suglasnosti potvrđuje da je taj postupak najbolji, dok ostali to ne potvrđuju. Kod naivnog Bayesovog klasifikatora po točnosti klasifikacije najbolji je postupak prilagodbe podataka odstranjivanjem šuma cenzure. Suprotno tome, težinska točnost klasifikacije i indeks suglasnosti sugeriraju da je za učenje naivnog Bayesovog klasifikatora najbolji postupak podvajanja cenzuriranih primjera. Postupak učenja regresijskog modela proporcionalnih hazarda iz diskretiziranih podataka na ovoj domeni nije konvergirao ($IBS \gg 0.25$), stoga je za usporedbu statističke sličnosti korišten model naučen iz izvornih (nediskretiziranih) podataka.

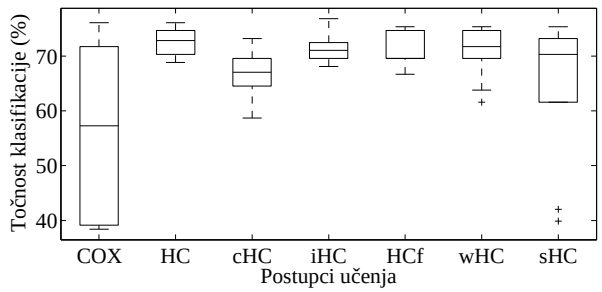
Statistički testovi predstavljeni su na slikama 7.6 (algoritam penjanja uzbrdo), 7.7 (algoritam uvjetnih neovisnosti) i 7.8 (naivni Bayesov klasifikator). Postupak učenja odstranjivanjem prekratko praćenih primjeraka za oba je algoritma učenja Bayesovih mreža (cHC, cCI) statistički lošiji od postupka podjele na vremenske intervale (iHC, iCI) i postupka učenja odstranjivanjem šuma cenzure (HCf, CIf). Postupak cHC je značajno lošiji i od postupka podvajanja cenzuriranih primjeraka (wHC). Ni jedan se drugi par postupaka sustavno (po većini metrika) statistički značajno ne razlikuje. Za naivni Bayesov klasifikator ne postoji suglasnost među statističkim testovima na korištenim metrikama.

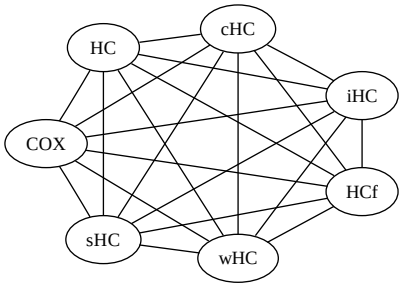
Tablica 7.3: Vrednovanje različitih postupaka učenja iz domene primarne bilijarne ciroze jetre. Korištene metrike su točnost klasifikacije, osjetljivost i specifičnost. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

	Mjera					
Postupak	Točnost klasifikacije (%)		Osjetljivost (%)		Specifičnost (%)	
Penjanje uzbrdo						
HC	72.14	(1.47)	53.33	(2.68)	84.79	(2.15)
cHC	68.59	(1.39)	30.99	(3.63)	**93.88**	(1.88)
iHC	**72.54**	(0.95)	48.20	(1.60)	88.91	(1.14)
HCf	71.38	(1.58)	70.45	(2.75)	72.00	(1.39)
wHC	69.86	(2.06)	81.44	(2.29)	62.06	(3.07)
sHC	69.49	(2.15)	**83.15**	(1.90)	60.30	(3.51)
Uvjetne neovisnosti						
CI	68.77	(1.54)	51.80	(3.76)	80.18	(2.25)
cCI	66.63	(2.16)	47.12	(4.18)	79.76	(1.92)
iCI	**69.31**	(1.91)	50.99	(3.58)	**81.64**	(1.51)
CIf	68.04	(1.83)	58.47	(4.81)	74.48	(2.67)
wCI	67.97	(2.14)	**71.80**	(3.68)	65.39	(2.05)
sCI	60.33	(1.93)	66.22	(3.01)	56.36	(2.65)
Naivni Bayesov klasifikator						
NB	77.14	(0.73)	63.78	(1.26)	86.12	(0.53)
cNB	76.59	(1.33)	62.52	(2.17)	86.06	(1.51)
iNB	77.07	(0.42)	63.06	(1.20)	**86.48**	(0.50)
NBf	**77.68**	(0.69)	75.86	(1.26)	78.91	(0.80)
wNB	71.45	(0.56)	**88.83**	(0.47)	59.76	(1.08)
Proporcionalni hazard						
COX	43.91	(1.64)	**95.14**	(1.91)	9.45	(1.67)
COXc	**72.97**	(0.77)	40.36	(1.46)	**94.91**	(0.71)

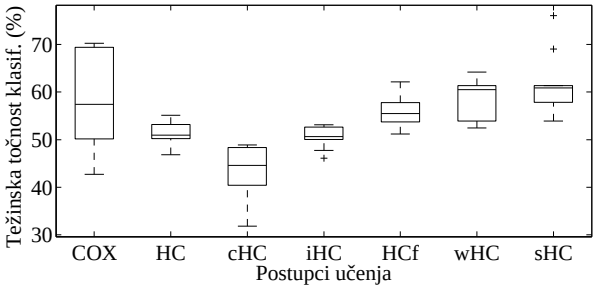
Tablica 7.4: Vrednovanje različitih postupaka učenja iz domene primarne bilijarne ciroze jetre. Korištene metrike su indeks suglasnosti, težinska točnost klasifikacije, integrirana Brierova ocjena i njena rezidualna varijacija. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

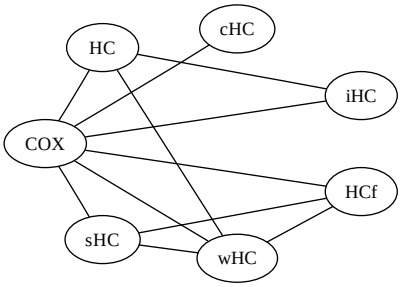
	Mjera							
Postupak	Težinska toč. klasif. (%)		Indeks suglasnosti		Integrirana Brierova ocjena		Rezidualna varijacija	
Penjanje uzbrdo								
HC	51.9	(0.9)	0.738	(0.014)	0.107	(0.001)	11.4	(0.6)
cHC	43.0	(1.3)	0.652	(0.021)	0.117	(0.002)	2.8	(1.1)
iHC	50.3	(0.7)	**0.768**	(0.010)	0.106	(0.001)	12.6	(0.6)
HCf	58.7	(1.1)	0.766	(0.017)	**0.103**	(0.002)	**14.4**	(1.2)
wHC	61.5	(0.7)	0.741	(0.015)	0.106	(0.001)	12.3	(0.8)
sHC	**62.2**	(0.9)	0.738	(0.014)	0.107	(0.001)	11.5	(0.5)
Uvjetne neovisnosti								
CI	50.9	(1.4)	0.722	(0.017)	0.111	(0.002)	8.0	(1.5)
cCI	48.0	(2.0)	0.691	(0.019)	0.113	(0.002)	6.5	(1.5)
iCI	51.3	(1.6)	**0.744**	(0.016)	0.110	(0.002)	9.2	(1.4)
CIf	53.7	(1.9)	0.702	(0.015)	0.111	(0.001)	8.0	(1.1)
wCI	**60.4**	(1.3)	0.739	(0.022)	**0.107**	(0.001)	**11.8**	(1.0)
sCI	57.2	(1.4)	0.675	(0.026)	0.112	(0.002)	7.3	(1.6)
Naivni Bayesov klasifikator								
NB	56.8	(0.5)	0.845	(0.008)	0.097	(0.001)	19.9	(0.3)
cNB	56.1	(1.1)	0.841	(0.008)	0.098	(0.001)	18.8	(0.3)
iNB	56.4	(0.4)	0.845	(0.009)	**0.097**	(0.001)	**20.0**	(0.2)
NBf	61.5	(0.5)	0.844	(0.006)	0.097	(0.001)	20.0	(0.4)
wNB	**66.7**	(0.3)	**0.849**	(0.007)	0.100	(0.001)	17.4	(0.2)
Proporcionalni hazard								
COX	**67.0**	(0.7)	0.086	(0.002)	0.768	(0.007)	-535.6	(7.1)
COXc	47.1	(0.7)	**0.818**	(0.011)	**0.078**	(0.001)	**34.8**	(1.2)



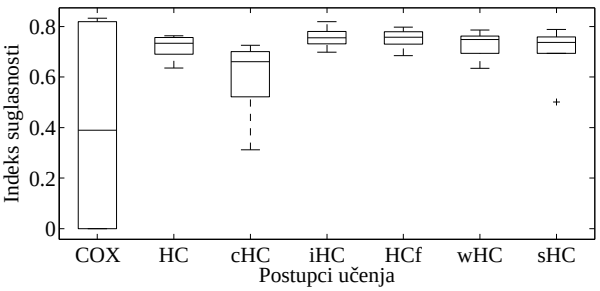


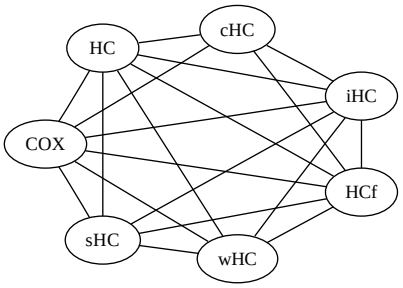
(a) Točnost klasifikacije



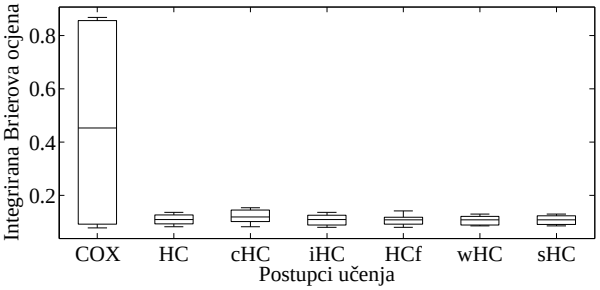


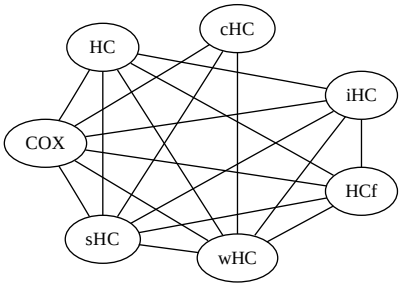
(b) Težinska točnost klasifikacije





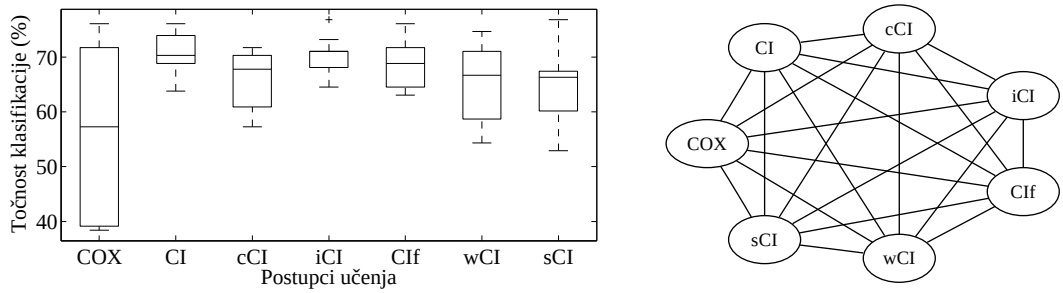
(c) Indeks suglasnosti



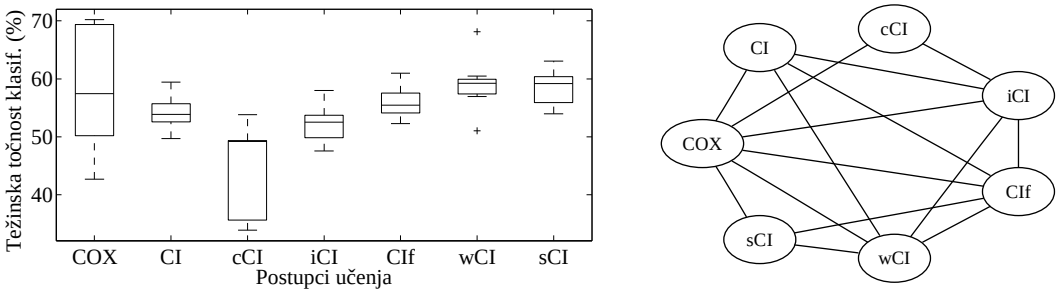


(d) Integrirana Brierova ocjena

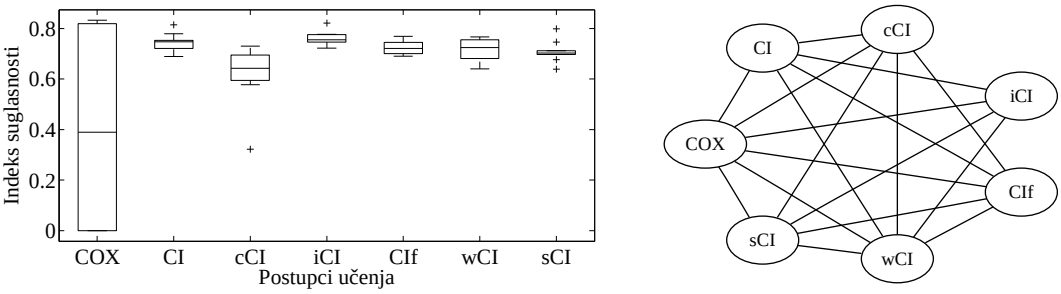
Slika 7.6: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena ciroze jetre).



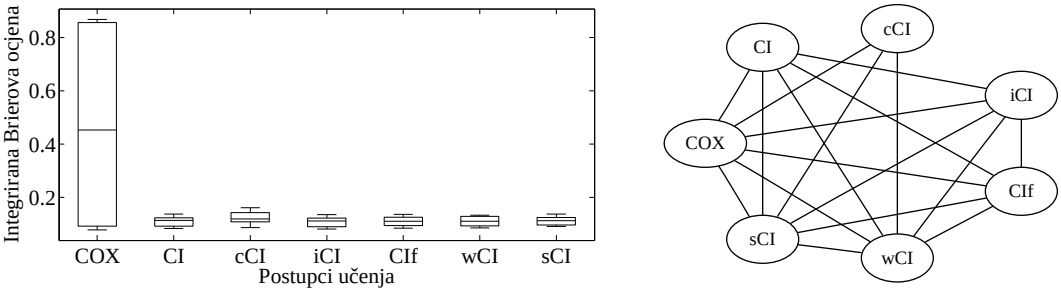
(a) Točnost klasifikacije



(b) Težinska točnost klasifikacije

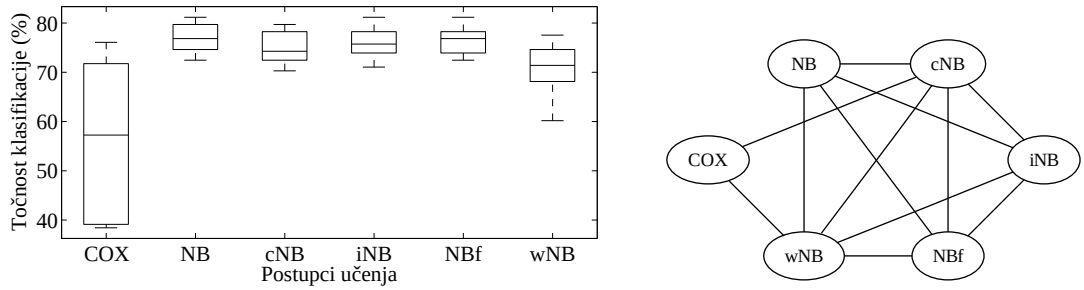


(c) Indeks suglasnosti

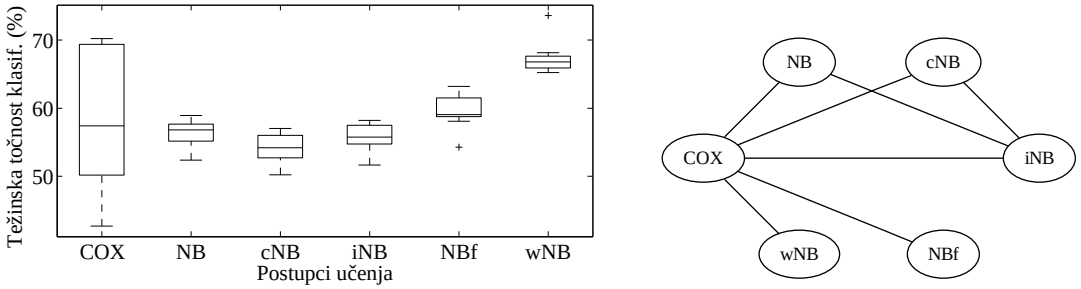


(d) Integrirana Brierova ocjena

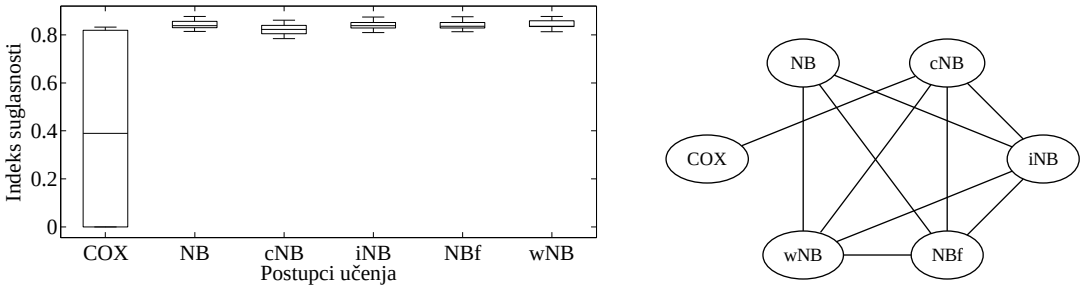
Slika 7.7: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena ciroze jetre).



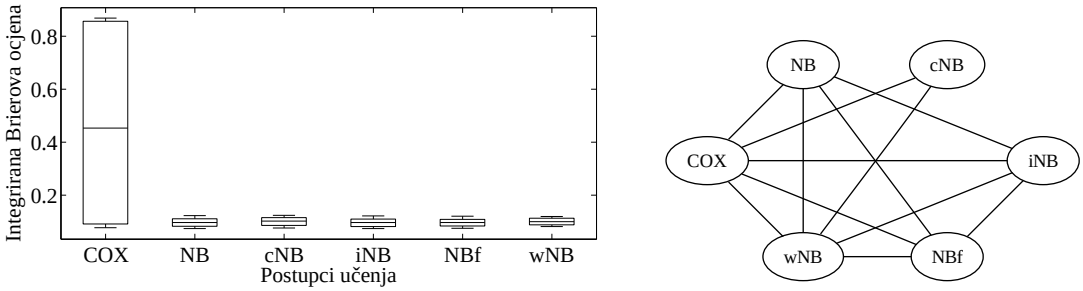
(a) Točnost klasifikacije



(b) Težinska točnost klasifikacije

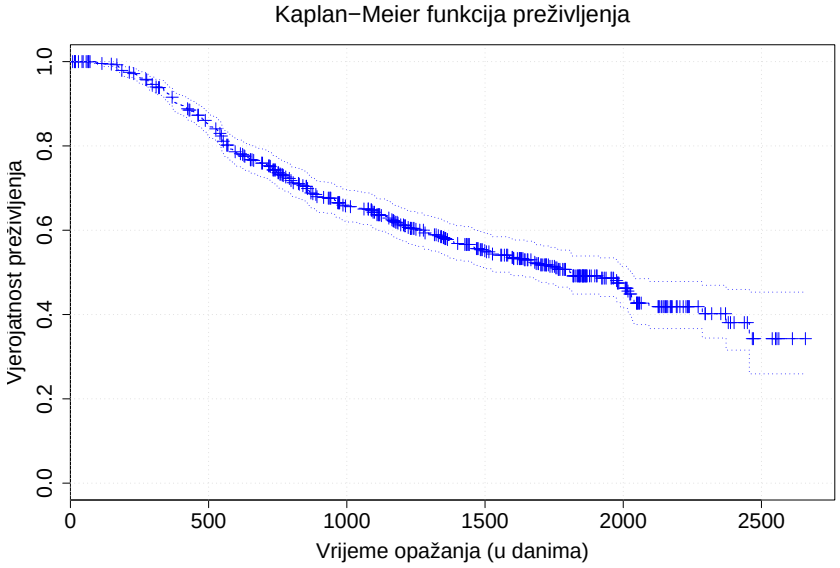


(c) Indeks suglasnosti



(d) Integrirana Brierova ocjena

Slika 7.8: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena ciroze jetre).



Slika 7.9: Funkcija preživljenja za domenu raka dojke. Funkcija je praćena krivuljama 95%-tnih intervala pouzdanosti (crtkano).

7.3 Rak dojke

Rak dojke nastaje kad normalne žljezdane stanice dojke promijene svoja svojstva te počnu nekontrolirano rasti, umnožavati se i uništavati okolno zdravo tkivo. Najčešći je zloćudni tumor u žena u svijetu.

Njemačka onkološka studijska skupina (engl. *German Breast Cancer Study Group 2*, GBSG2) proučavala je načine liječenja raka dojke s pozitivnim limfnim čvorovima (Schumacher *et al.*, 1994). Studija je bila provedena nad 686 žena starosti do zaključno 65 godina, koje su imale pozitivne regionalne limfne čvorove bez udaljenih metastaza. Promatrao se moguć povrat raka dojke kod pacijentica. Funkcija preživljenja predstavljena je na slici 7.9. Domena je dostupna kao dio softverskog statističkog paketa *ipred* kao dijela okružja *R* (R Development Core Team, 2008). Također je dostupna i na adresi `http://www.blackwellpublishing.com/rss/Volumes/A162p1.htm` (dostupno: 21. prosinca 2009).

Svaki je zapis o pacijenticama sastavljen iz 7 prognostičkih kovarijata, pokazateljem hormonalne terapije, vremenom promatranja i statusom. Status označava je li promatranje pacijentice bilo cenzurirano ili je završilo recidivom raka dojke. Diskretizirane su sljedeće kovarijate: starost pacijentice, veličina tumora, broj pozitivnih limfnih čvorova, razina progesterona i razina estrogena.

Odabrane granice za postupak prilagodbe podataka podjelom na vremenske intervale su $(0, 500, 1000, 1500, 2000, \infty)$. Rezultati stratificirane deseterostruke unakrsne provjere na deset iteracija predstavljeni su u tablici 7.5 (standardne metrike) i tablici 7.6 (metrike analize preživljenja). Postupak prilagodbe podataka podjelom na vremen-

ske intervale po točnosti klasifikacije je na samom vrhu, za sva tri postupka učenja. Tu tezu potvrđuju indeks suglasnosti i integrirana Brierova ocjena. Težinska točnost klasifikacije sugerira pak, da je najbolji postupak podvajanja cenzuriranih primjeraka.

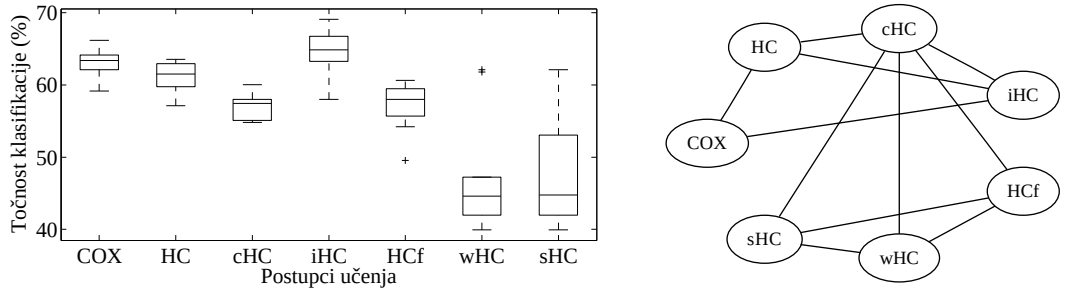
Statistički testovi predstavljeni su na slikama 7.10 (algoritam penjanja uzbrdo), 7.11 (algoritam uvjetnih neovisnosti) i 7.12 (naivni Bayesov klasifikator). I indeks suglasnosti, i integrirana Brierova ocjena sugeriraju da se za algoritam HC postupak podjele na vremenske intervale statistički značajno ne razlikuje od Coxove regresije i postupka odstranjivanja šuma cenzure (HCf). Za algoritam CI i NB, statistički su podjednako dobri postupak podjele na vremenske intervale, na jednoj strani, te Coxova regresija i postupak podvajanja cenzuriranih primjeraka (sCI, wNB) na drugoj. Tretiranje cenzuriranih primjeraka kao negativnih za naivni Bayesov klasifikator je također statistički podjednako postupku iNB.

Tablica 7.5: Vrednovanje različitih postupaka učenja iz domene raka dojke. Korištene metrike su točnost klasifikacije, osjetljivost i specifičnost. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

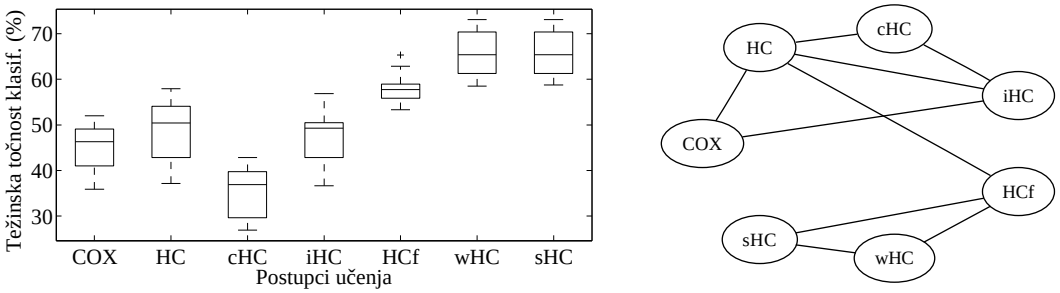
	Mjera					
Postupak	Točnost klasifikacije (%)		Osjetljivost (%)		Specifičnost (%)	
Penjanje uzbrdo						
HC	59.33	(0.64)	35.38	(3.60)	77.83	(3.64)
cHC	56.41	(0.00)	0.00	(0.00)	**100.00**	(0.00)
iHC	**64.83**	(0.51)	26.25	(1.93)	94.63	(0.83)
HCf	56.75	(0.66)	75.52	(2.43)	42.25	(2.37)
wHC	44.66	(0.96)	95.38	(2.95)	5.48	(3.60)
sHC	44.15	(0.71)	**97.49**	(2.38)	2.95	(2.87)
Uvjetne neovisnosti						
CI	61.87	(0.72)	23.38	(2.10)	91.60	(2.15)
cCI	56.37	(0.12)	0.03	(0.11)	**99.90**	(0.18)
iCI	**63.21**	(0.71)	21.44	(1.84)	95.48	(0.94)
CIf	57.07	(0.73)	72.34	(1.78)	45.27	(1.53)
wCI	44.68	(0.91)	96.79	(2.00)	4.42	(1.87)
sCI	43.62	(0.35)	**98.33**	(2.01)	1.34	(1.55)
Naivni Bayesov klasifikator						
NB	62.01	(0.41)	53.51	(0.72)	68.58	(0.49)
cNB	58.54	(0.50)	6.02	(1.01)	**99.12**	(0.39)
iNB	**68.53**	(0.57)	43.58	(1.12)	87.80	(0.65)
NBf	58.41	(0.39)	76.89	(1.41)	44.13	(1.53)
wNB	52.65	(0.48)	**91.24**	(1.49)	22.84	(1.03)
Proporcionalni hazard						
COX	**63.16**	(0.28)	**26.09**	(0.42)	91.81	(0.50)
COXc	61.79	(0.43)	22.88	(0.89)	**91.86**	(0.33)

Tablica 7.6: Vrednovanje različitih postupaka učenja iz domene raka dojke. Korištene metrike su indeks suglasnosti, težinska točnost klasifikacije, integrirana Brierova ocjena i njena rezidualna varijacija. Istaknuti su najbolji rezultati unutar svake skupine strojnog učenja.

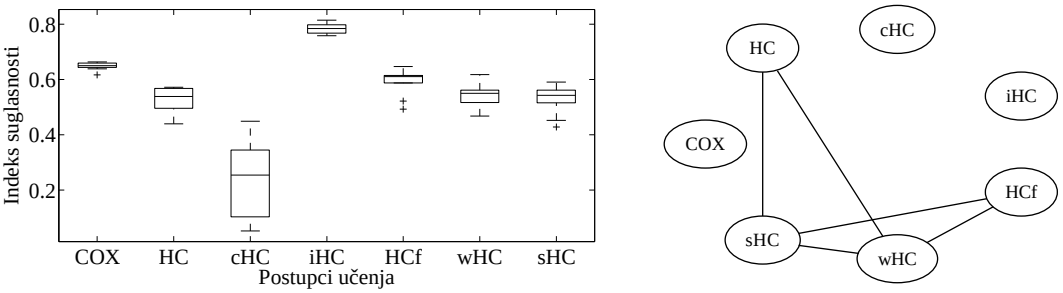
	Mjera							
Postupak	Težinska toč. klasif. (%)		Indeks suglasnosti		Integrirana Brierova ocjena		Rezidualna varijacija	
Penjanje uzbrdo								
HC	47.9	(1.4)	0.559	(0.010)	0.122	(0.000)	3.2	(0.3)
cHC	34.2	(0.1)	0.208	(0.039)	0.143	(0.001)	-13.9	(0.3)
iHC	46.2	(0.9)	**0.781**	(0.007)	**0.118**	(0.001)	**5.8**	(0.4)
HCf	61.2	(0.9)	0.617	(0.007)	0.120	(0.001)	4.5	(0.7)
wHC	64.8	(0.8)	0.567	(0.009)	0.118	(0.000)	5.6	(0.2)
sHC	**65.3**	(0.5)	0.564	(0.014)	0.119	(0.000)	5.6	(0.1)
Uvjetne neovisnosti								
CI	43.9	(0.7)	0.588	(0.010)	0.122	(0.001)	2.9	(0.4)
cCI	34.2	(0.1)	0.225	(0.054)	0.143	(0.001)	-14.1	(0.3)
iCI	43.7	(0.8)	**0.780**	(0.009)	**0.118**	(0.001)	**5.7**	(0.3)
CIf	60.1	(0.8)	0.550	(0.008)	0.125	(0.002)	0.7	(1.3)
wCI	65.2	(0.7)	0.549	(0.012)	0.121	(0.001)	3.8	(0.3)
sCI	**65.2**	(0.8)	0.592	(0.012)	0.119	(0.001)	5.1	(0.3)
Naivni Bayesov klasifikator								
NB	54.7	(0.3)	0.657	(0.005)	0.119	(0.000)	5.4	(0.3)
cNB	36.7	(0.5)	0.639	(0.005)	0.135	(0.000)	-7.3	(0.3)
iNB	53.6	(0.5)	**0.757**	(0.004)	**0.115**	(0.000)	**8.4**	(0.3)
NBf	61.8	(0.5)	0.646	(0.005)	0.116	(0.001)	7.5	(0.5)
wNB	**65.5**	(0.6)	0.655	(0.005)	0.117	(0.000)	7.1	(0.1)
Proporcionalni hazard								
COX	**45.2**	(0.2)	0.652	(0.004)	**0.110**	(0.000)	**12.5**	(0.2)
COXc	43.9	(0.4)	**0.677**	(0.004)	0.114	(0.000)	9.7	(0.2)



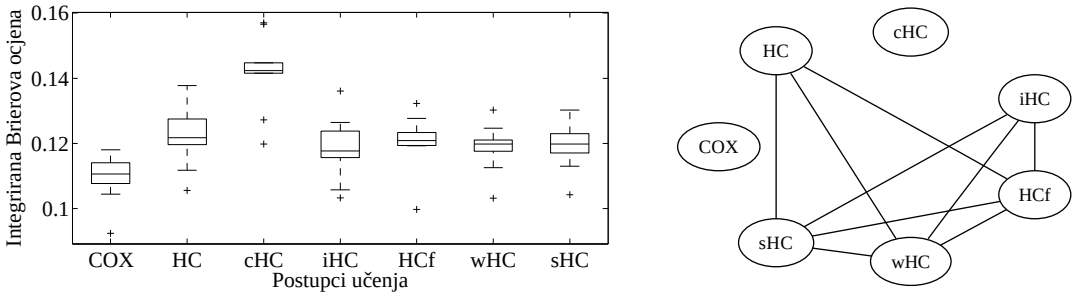
(a) Točnost klasifikacije



(b) Težinska točnost klasifikacije

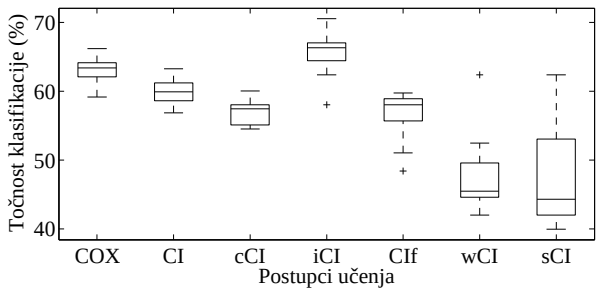


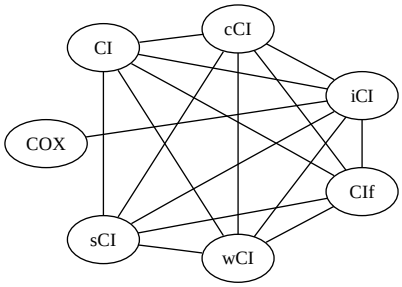
(c) Indeks suglasnosti



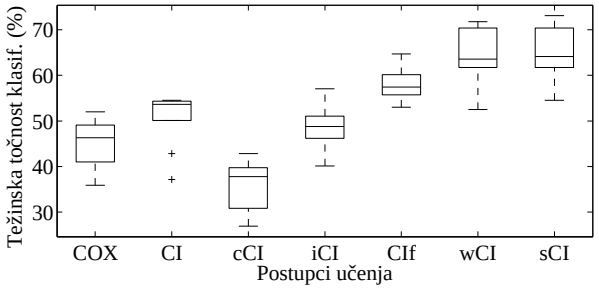
(d) Integrirana Brierova ocjena

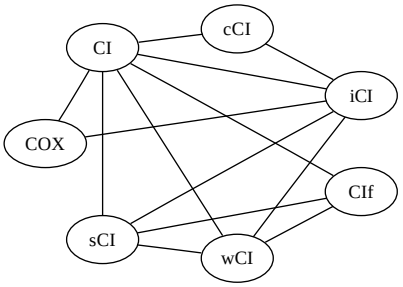
Slika 7.10: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena raka dojke).



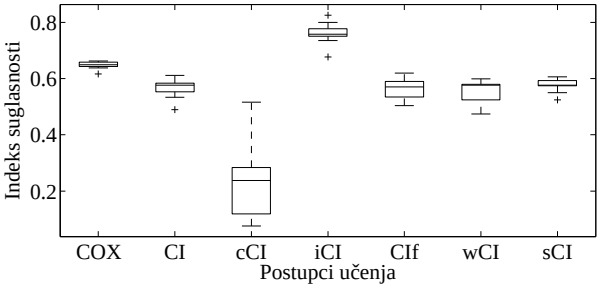


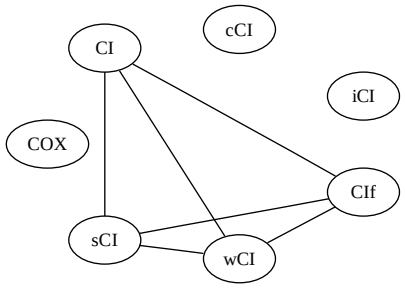
(a) Točnost klasifikacije



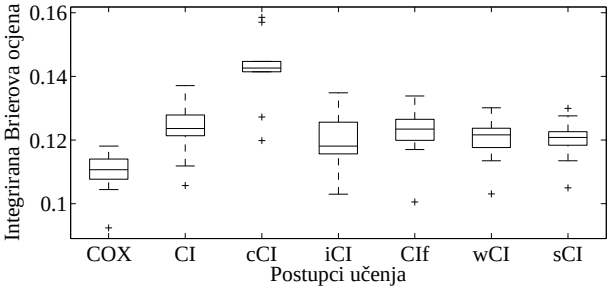


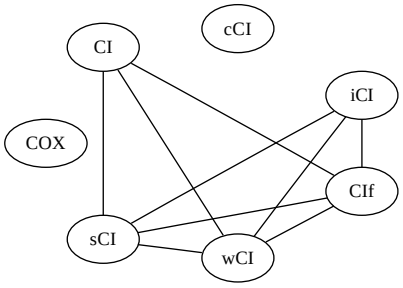
(b) Težinska točnost klasifikacije





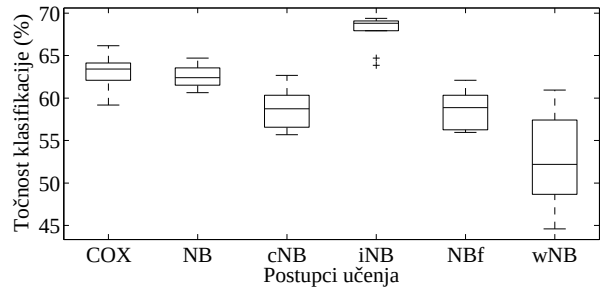
(c) Indeks suglasnosti

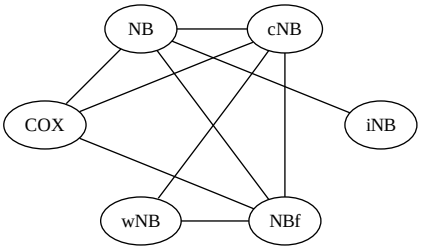




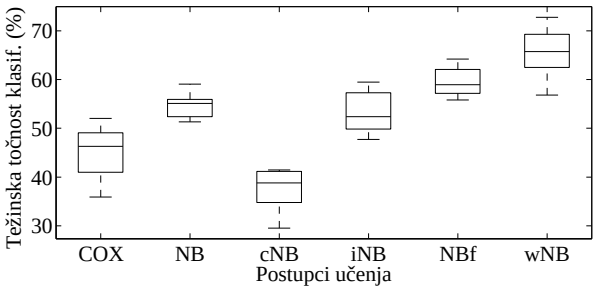
(d) Integrirana Brierova ocjena

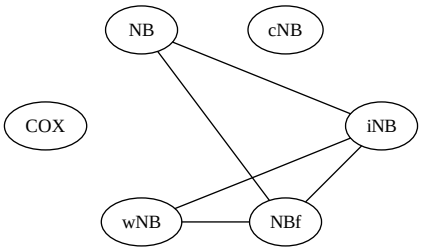
Slika 7.11: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena raka dojke).



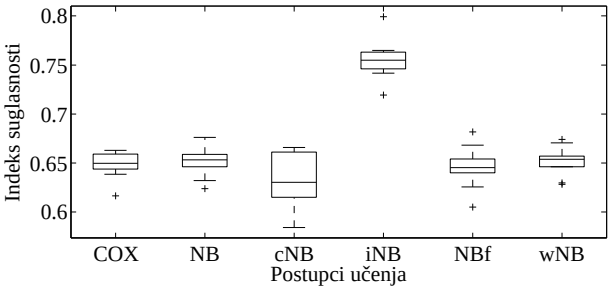


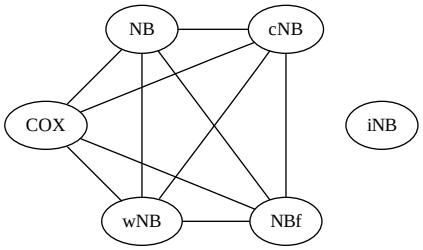
(a) Točnost klasifikacije



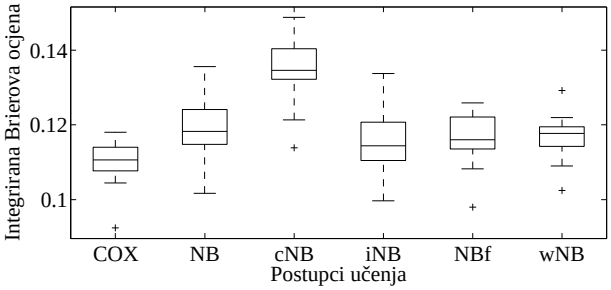


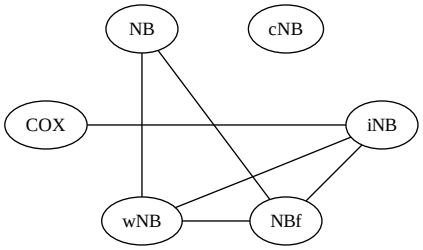
(b) Težinska točnost klasifikacije





(c) Indeks suglasnosti





(d) Integrirana Brierova ocjena

Slika 7.12: Box i Whisker dijagrami 5 x 2 testa predstavljenih postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda za različite metrike vrednovanja, te njima pridruženi dijagrami statističke sličnosti (domena raka dojke).

7.4 Rezultati

Rezultati testova na realnim domenama ne pružaju jedinstven odgovor na pitanje o tome koji je postupak pripremne obrade cenzuriranih podataka najbolji za potrebe strojnog učenja. Standardna točnost klasifikacije sugerira da je postupak učenja podjelom na vremenske intervale u prosjeku najbolji, no njena je važnost upitna zbog udjela cenzure u podacima. Metrike analize preživljenja istovremeno sugeriraju da su pretežno (po broju "pobjeda") najbolji postupci učenja podvajanjem cenzuriranih primjeraka uz težinske faktore i postupak učenja podjelom na vremenske intervale. Statističko vrednovanje rezultata testova na realnim domenama slijedi nekonzistentnost vrijednosti dobivenih različitim metrikama te ne izolira niti jedan postupak kao najbolji ili najgori. To je u velikoj mjeri uzrokovano smanjivanjem α vrijednosti Bonferronijevom korekcijom uslijed velikog broja testova hipoteze. Statistički testovi značajno ($\alpha = 0.05$) ne razlikuju postupke podvajanja cenzuriranih primjeraka uz težinske faktore i postupak prilagodbe podataka odstranjivanjem šuma cenzure.

Poglavlje 8

Zaključak

Jedan od ciljeva ove disertacije bio je temeljito usporediti postupke rukovanja cenzuriranim podacima o preživljenju za potrebe strojnog učenja, s naglaskom na učenje Bayesovih mreža. Uspoređeni su sljedeći postojeći postupci: tretiranje cenzuriranih primjeraka kao negativnih, odstranjivanje prekratko praćenih primjeraka, učenje podjelom na vremenske intervale te učenje podvajanjem cenzuriranih primjeraka uz težinske faktore; predložen je i jedan nov postupak, rukovanje cenzuriranim podacima o preživljenju odstranjivanjem šuma cenzure. Postupci su bili primijenjeni na učenje Bayesovih mreža dvama, dobro poznatim algoritmima: 1) algoritmom penjanja uzbrdo, kao predstavnikom algoritama temeljenih na uspjehu i 2) algoritmom uvjetnih neovisnosti, kao predstavnikom algoritama temeljenih na neovisnostima. Dodatno su, radi kvalitetnije usporedbe, bili primijenjeni i na učenje naivnog Bayesovog klasifikatora, kao jednog od najpoznatijih i najučinkovitijih modela predstavljanja znanja u strojnom učenju. S obzirom na to da je problem učenja iz cenzuriranih podataka o preživljenju proizašao iz analize preživljenja, svi su postupci dodatno uspoređeni i s Coxovom regresijom.

Kako bi se provjerila i usporedila sposobnost navedenih postupaka rukovanja cenzuriranim podacima o preživljenju, provedeno je temeljito testiranje karakteristika naučenih modela. U simulacijskoj studiji slučajno je uzorkovano 100 skupova podataka iz modela određenog korelacijskom matricom i regresijskim koeficijentima. Svaki je skup podataka bio slučajno cenzuriran s nekoliko razina cenzure (od 0% do 80%). Vrsnost naučenih modela vrednovala se, kako standardnim metrikama strojnog učenja (točnost klasifikacije, osjetljivost, specifičnost), tako i metrikama analize preživljenja (težinska točnost klasifikacije, indeks suglasnosti, integrirana Brierova ocjena). S obzirom na to da su se u simulacijskoj studiji za testiranje standardne točnosti klasifikacije koristili stvarni, to jest necenzurirani ishodi, ona je uzeta kao temeljna ispravna mjera za usporedbu kvalitete ostalih metrika, onih analize preživljenja. Za srednju razinu cenzure (od 20% do približno 50%), najvišu točnost klasifikacije ima postupak podvajanja cenzuriranih primjeraka uz težinske faktore. Za visoku razinu cenzure (od približno 50% do 80%), najvišu točnost klasifikacije ima postupak prilagodbe podataka odstranjivanjem šuma cenzure. On jedini nadmašuje Coxovu regresiju na svim razinama cenzure. Statistička usporedba srednjih rangova dobivenih točnosti klasifikacije uglavnom potvrđuje

ove odnose uz razinu statističke značajnosti testa $\alpha = 0.05$. Evidentan je rast performansi Coxove regresije i postupka prilagodbe podataka odstranjivanjem šuma cenzure (za sve modele) s rastom udjela cenzure u podacima. Dobivene težinske točnosti klasifikacija najbliže prate relativan odnos dobivenih standardnih točnosti klasifikacija. Iako sugerira da je Coxova regresija za sve razine cenzure najbolja, od ostalih postupaka, na razini cenzure do 50%, indeks suglasnosti ističe postupak podjele na vremenske intervale. Integrirana Brierova ocjena za razine cenzure do približno 60% kao najbolju ističe Coxovu regresiju, a za više razine cenzure, pak postupak prilagodbe podataka odstranjivanjem šuma cenzure.

Sposobnost postupaka rukovanja cenzuriranim podacima o preživljenju, za ispravno učenje topologija Bayesovih mreža, testirana je u simulacijskoj studiji. U njoj je, uz neka ograničenja, slučajno generirano ukupno 100 ishodišnih modela nalik modelima očekivanim u kliničkoj medicini. Ispravnost učenja topologija mreža mjerila se brojem dodanih, manjkajućih i obrnuto usmjerenih lukova, suma kojih predstavlja broj potrebnih elementarnih promjena na mreži, kako bi se ona dovela u ishodišni oblik. U prosjeku su najbolji postupak podvajanja cenzuriranih primjeraka uz težinske faktore i tretiranje cenzuriranih primjeraka kao negativnih. Nešto je lošiji postupak prilagodbe podataka odstranjivanjem šuma cenzure, jer u prosjeku doda više lukova razredu. Iznenađujuće, gledajući broj potrebnih elementarnih promjena na cijeloj mreži, na gotovo svim razinama cenzure ne postoji statistički značajna razlika ($\alpha = 0.05$) između navedena tri postupka. Gledajući broj potrebnih elementarnih promjena na lukovima vezanim uz razred, samo se prva dva postupka statistički značajno ne razlikuju.

Za testiranje rada opisanih postupaka na realnim domenama bilo je potrebno koristiti više različitih metrika vrednovanja. Usporedbom isključivo rezultata točnosti klasifikacije bili bismo navedeni na krivi put zbog utjecaja cenzure. Standardna točnost klasifikacije sugerira da je postupak učenja podjelom na vremenske intervale u prosjeku najbolji. Metrike analize preživljenja istovremeno sugeriraju da je uz taj postupak, podjednako dobar i postupak učenja podvajanjem cenzuriranih primjeraka uz težinske faktore. Statističko vrednovanje postupaka učenja na realnim domenama, pokazalo je da se oba postupka podvajanja cenzuriranih primjeraka (uz težinske faktore) i postupak prilagodbe podataka odstranjivanjem šuma cenzure, statistički značajno ne razlikuju. Pojedinačno se pak na nekim domenama statistički značajno razlikuju ($\alpha = 0.05$) od pojedinačnih drugih postupaka, no niti jedan od njih nije na svim domenama statistički značajno različit od nekog drugog postupka.

Na osnovi predstavljenih rezultata može se zaključiti, kako je predloženi postupak prilagodbe podataka odstranjivanjem šuma cenzure, sa svrhom učenja točnih modela klasifikacije iz cenzuriranih podataka o preživljenju, izvrsno rješenje. Na to ukazuju izvrsni, statistički poduprijeti rezultati, postignuti u simulacijskoj studiji. Istovremeno, unatoč nešto lošijim performansama na realnim domenama, statističkim je testovima dokazano da se predloženi postupak statistički značajno ne razlikuje ($\alpha = 0.05$) od postupaka učenja podvajanjem cenzuriranih primjeraka uz težinske faktore, koji su se na metrikama analize preživljenja pokazali kao pretežno najbolji (uz postupak podjele na vremenske intervale). Što se tiče ispravnog otkrivanja strukture Bayesovih mreža,

predložen postupak daje nešto kompleksnije topologije lukova vezanih uz razred, ali se na razini broja potrebnih elementarnih promjena na cijeloj mreži, statistički značajno ne razlikuje ($\alpha = 0.05$) od najuspješnijih postupaka.

Dodaci

Dodatak A

Podaci iz domene koronarne bolesti srca

U nastavku su detaljnije predstavljeni podaci iz domene koronarne bolesti srca, korišteni za ilustraciju rada Bayesovih mreža i osnovnih koncepata analize preživljenja (poglavlje 2).

Podaci predstavljeni u tablici A.1 generirani su iz Bayesove mreže, određene topologijom sa slike 2.2 i tablicama uvjetnih vjerojatnosti sa slike 2.3, postupkom uzorkovanja, opisanim u sekciji 6.2.2. Vremena preživljenja generirana su iz eksponencijalne distribucije (sekcija 6.1.3) uz proizvoljno određen vektor smislenih regresijskih koeficijenata $\beta = (0.4, -0.3, -0.3, 0.1, 0, 0.7, -0.5)$ pridružen vektoru kovarijata $\mathbf{X} = (X_P, X_C, X_S, X_T, X_A, X_E, X_U)$. Oblik krivulje distribucije vremena preživljenja određen je koeficijentom $\lambda = 0.07$. Tablica A.2 predstavlja iste primjerke po uvođenju približno 60%-tne umjetne cenzure. Oblik krivulje distribucije vremena praćenja određen je koeficijentom $\lambda = 0.2$.

Tablica A.1: Umjetno generirani podaci za domenu koronarne bolesti srca. Zadnji stupac (T) predstavlja vrijeme preživljenja.

Prehrana	*Cigarete*	*Stres*	*Tlak*	*Anemija*	*EKG*	*Umor*	*KBS*	T
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	23
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	50
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Prisutan	Odsutan	3
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	44
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	8
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	7
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	9
Loša	Da	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	6
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	39
Dobra	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	3
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	8
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	31
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	18
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	14
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Da	Prisutan	Normalan	Odsutna	Normalan	Prisutan	Odsutan	17
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	6
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Odsutan	Odsutan	15
Loša	Da	Prisutan	Povišen	Odsutna	Normalan	Prisutan	Prisutan	4
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	7
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	43
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	6
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	7
Dobra	Da	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	7
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	8
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Prisutan	Prisutan	9
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	14
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	4
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	15
Loša	Ne	Prisutan	Povišen	Prisutna	Normalan	Prisutan	Prisutan	2
Loša	Da	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	11
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Prisutna	Normalan	Prisutan	Odsutan	3
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Prisutan	Prisutan	9
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	4
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	14
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	48
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	25
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	51
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	7
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	9
Loša	Da	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	6

Prehrana	*Cigarete*	*Stres*	*Tlak*	*Anemija*	*EKG*	*Umor*	*KBS*	*T*
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	39
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	2
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	4
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	10
Dobra	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	6
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	43
Dobra	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	5
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	5
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	9
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	4
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	4
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	30
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	2
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	6
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	48
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	52
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	4
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	3
Dobra	Da	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	1
Loša	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	4
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	42
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	41
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Prisutan	11
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	5
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	29
Dobra	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	3
Dobra	Da	Prisutan	Normalan	Prisutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	2
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	3
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	5
Loša	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Ne	Prisutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	8
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Da	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	36
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	8
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	24
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	2
Dobra	Da	Odsutan	Normalan	Prisutna	Normalan	Prisutan	Odsutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	53
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	9
Dobra	Da	Odsutan	Normalan	Prisutna	Normalan	Prisutan	Odsutan	16
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	11
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	6
Loša	Ne	Odsutan	Normalan	Prisutna	Normalan	Odsutan	Odsutan	29

Tablica A.2: Umjetno generirani podaci za domenu koronarne bolesti srca nakon uvođenja približno 60%-tne umjetne cenzure. Zadnji stupac (T) predstavlja vrijeme praćenja.

Prehrana	*Cigarete*	*Stres*	*Tlak*	*Anemija*	*EKG*	*Umor*	*KBS*	T
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	8
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	48
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Prisutan	Odsutan	2
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	14
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	4
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	2
Loša	Da	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	3
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	14
Dobra	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	2
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	6
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	2
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Da	Prisutan	Normalan	Odsutna	Normalan	Prisutan	Odsutan	12
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	5
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Odsutan	Odsutan	15
Loša	Da	Prisutan	Povišen	Odsutna	Normalan	Prisutan	Prisutan	4
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	1
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	43
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	6
Dobra	Da	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	3
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	8
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	3
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	15
Loša	Ne	Prisutan	Povišen	Prisutna	Normalan	Prisutan	Prisutan	2
Loša	Da	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	5
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Odsutan	Povišen	Prisutna	Normalan	Prisutan	Odsutan	3
Loša	Ne	Prisutan	Povišen	Prisutna	Abnormalan	Prisutan	Odsutan	7
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	4
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	4
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	25
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	10
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	9
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	2
Loša	Da	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	3

Prehrana	*Cigarete*	*Stres*	*Tlak*	*Anemija*	*EKG*	*Umor*	*KBS*	*T*
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	4
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	2
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	3
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	2
Dobra	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	9
Dobra	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	1
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	3
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	2
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	12
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	2
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	1
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	3
Dobra	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	4
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	2
Dobra	Da	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	1
Loša	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	2
Dobra	Da	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	34
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Prisutan	11
Dobra	Ne	Odsutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	7
Dobra	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	1
Loša	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	3
Dobra	Da	Prisutan	Normalan	Prisutna	Abnormalan	Prisutan	Prisutan	2
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Prisutan	2
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	1
Dobra	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	2
Loša	Da	Odsutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	3
Dobra	Ne	Prisutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	8
Dobra	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	2
Dobra	Da	Odsutan	Normalan	Odsutna	Abnormalan	Odsutan	Odsutan	6
Loša	Ne	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Odsutan	1
Dobra	Ne	Prisutan	Povišen	Odsutna	Normalan	Odsutan	Odsutan	9
Loša	Ne	Odsutan	Normalan	Odsutna	Abnormalan	Prisutan	Prisutan	2
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Prisutan	Prisutan	2
Dobra	Da	Odsutan	Normalan	Prisutna	Normalan	Prisutan	Odsutan	2
Loša	Ne	Prisutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	9
Loša	Da	Prisutan	Povišen	Odsutna	Abnormalan	Odsutan	Odsutan	2
Dobra	Da	Odsutan	Normalan	Prisutna	Normalan	Prisutan	Odsutan	16
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	1
Loša	Ne	Odsutan	Normalan	Odsutna	Normalan	Odsutan	Odsutan	3
Loša	Ne	Odsutan	Normalan	Prisutna	Normalan	Odsutan	Odsutan	17

Dodatak B

Neki koncepti iz teorije vjerojatnosti

U nastavku su predstavljeni temeljni koncepti teorije vjerojatnosti nužni za razumijevanje mehanizma Bayesovih mreža. U disertaciji su se za rad s Bayesovim mrežama koristile isključivo diskretne varijate te su samo one predstavljene u dodatku. Detaljniji pregled relevantnih koncepata teorije vjerojatnosti može se pronaći u Alpaydin (2004); Berthold & Hand (2003); Bishop (2007); Borgelt & Kruse (2002), odakle su i preuzeti opisi združene, uvjetne i potpune vjerojatnosti.

B.1 Vjerojatnost

Ukoliko su sve vrijednosti $(a_1, \ldots, a_r)$, koje pretpostavljena slučajna varijata A može poprimiti jednako izgledne *a priori*, vjerojatnost ishoda a_i može se odrediti kao broj pojavljivanja tog ishoda u ukupnom broju neovisnih pokusa $l = \sum_j \#a_j$ (frekventistička teorija), odnosno:

$$\mathrm{P}(A = a_i) = \frac{\#a_i}{l}. \tag{B.1}$$

Radi veće preglednosti i jednostavnosti, navedena se vjerojatnost $\mathrm{P}(A = a_i)$ obično obilježava s $\mathrm{P}(A_i)$. Ukoliko nas umjesto vjerojatnosti konkretnog ishoda zanima distribucija vjerojatnosti (iz koje se lako dolazi do pojedinih vjerojatnosti), tada se koristi notacija $\mathrm{P}(A)$.

B.2 Združena vjerojatnost

Vjerojatnost da varijata A ima vrijednost a_i i da varijata B ima vrijednost b_j zove se združena vjerojatnost (engl. *joint probability*) i označava se s $\mathrm{P}(A = a_i, B = b_j)$. Određena je brojem preklapanja ishoda a_i i b_j u ukupnom broju neovisnih pokusa l:

$$\mathrm{P}(A = a_i, B = b_j) = \frac{\#(a_i \wedge b_j)}{l}. \tag{B.2}$$

Ukoliko su slučajne varijate A i B međusobno neovisne, njihova je združena vjerojatnost jednaka produktu njihovih zasebnih vjerojatnosti:

$$\mathrm{P}(A, B) = \mathrm{P}(A) \cdot \mathrm{P}(B). \tag{B.3}$$

B.3 Uvjetna vjerojatnost

U slučaju da su varijate A i B međusobno ovisne, vjerojatnost ishoda varijate A, uz zabilježen ishod varijate B, zove se uvjetna vjerojatnost A, u odnosu na B (engl. *conditional probability*), i označava s $\mathrm{P}(A|B)$. Određena je izrazom:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)}. \tag{B.4}$$

Zbog komutativnosti operacije, združena se distribucija vjerojatnosti u ovom slučaju može izraziti na sljedeći način:

$$\mathrm{P}(A, B) = \mathrm{P}(A|B) \cdot \mathrm{P}(B) = \mathrm{P}(B|A) \cdot P(A). \tag{B.5}$$

Iz izraza B.5 slijedi Bayesova formula:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \cdot \mathrm{P}(A)}{\mathrm{P}(B)}. \tag{B.6}$$

B.4 Potpuna vjerojatnost

Vjerojatnost ishoda slučajne varijate može se izračunati kao težinski prosjek svih uvjetnih vjerojatnosti u odnosu na ostale varijate, korištenjem zakona potpune vjerojatnosti (engl. *total probability*). Za slučajnu varijatu A u odnosu na slučajnu varijatu B, primjenom zakona potpune vjerojatnosti, dolazimo do izraza:

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i). \tag{B.7}$$

Ovaj je postupak poznat i pod nazivom marginaliziranje (engl. *marginalization*), u ovom slučaju po varijati B.

B.5 Pravilo lanca

U teoriji vjerojatnosti pravilo lanca omogućuje određivanje vrijednosti bilo kojeg člana združene distribucije vjerojatnosti skupa slučajnih varijata korištenjem uvjetnih vjerojatnosti. Skup varijata $X_1, \ldots, X_n$ združenih distribucijom vjerojatnosti može se odrediti izrazom:

$$\begin{aligned} \mathrm{P}(X_1, \ldots, X_n) &= \mathrm{P}(X_n | X_{n-1}, \ldots, X_1) \cdot \mathrm{P}(X_{n-1}, \ldots, X_1) \\ &= \prod_{i=1}^{n} \mathrm{P}(X_i | X_{i-1}, \ldots, X_1). \end{aligned} \tag{B.8}$$

Primjenom na Bayesove mreže izraz poprima oblik:

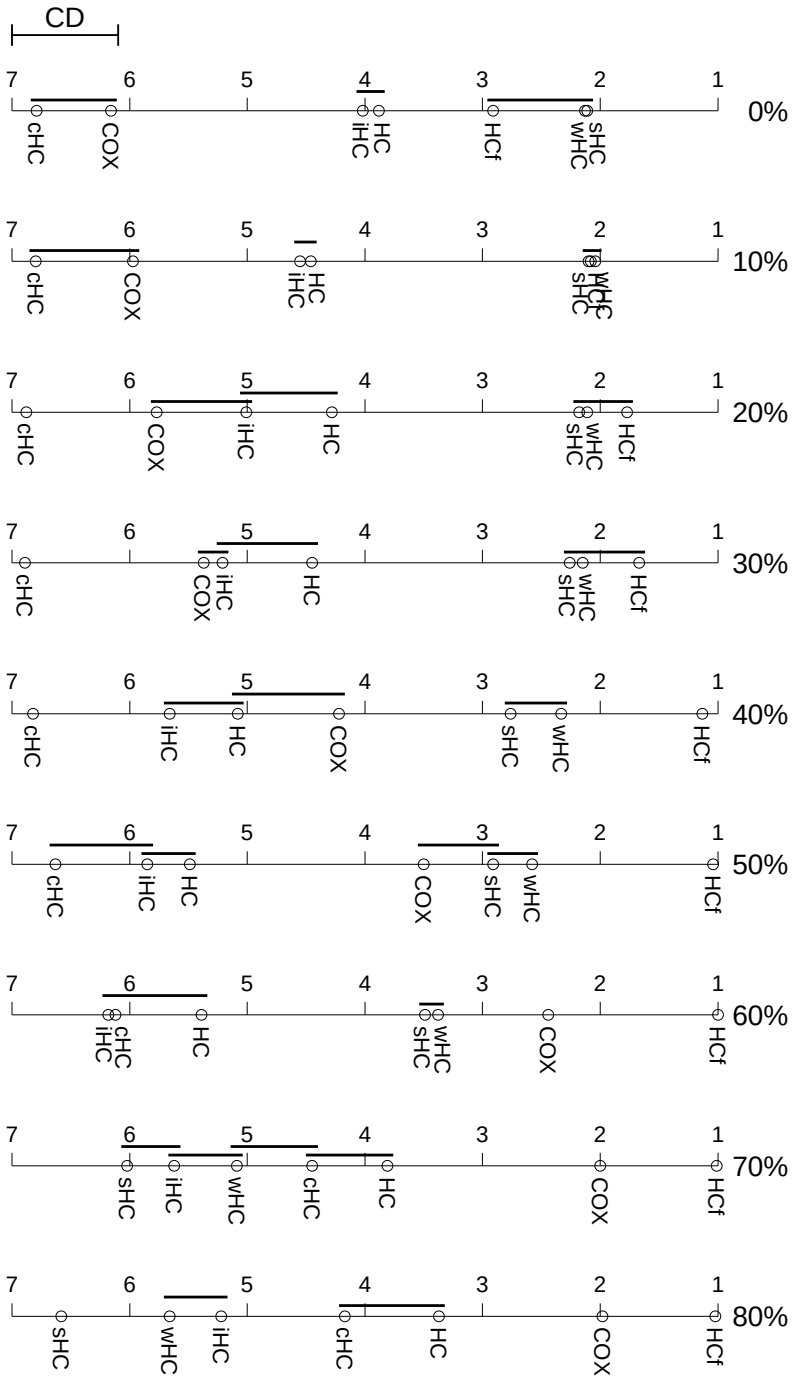
$$\mathrm{P}(V_1, \ldots, V_n) = \prod_{i=1}^{n} \mathrm{P}(V_i | \boldsymbol{\pi}(V_i)), \tag{B.9}$$

gdje je $\boldsymbol{\pi}(V_i)$ skup roditelja čvorišta V_i.

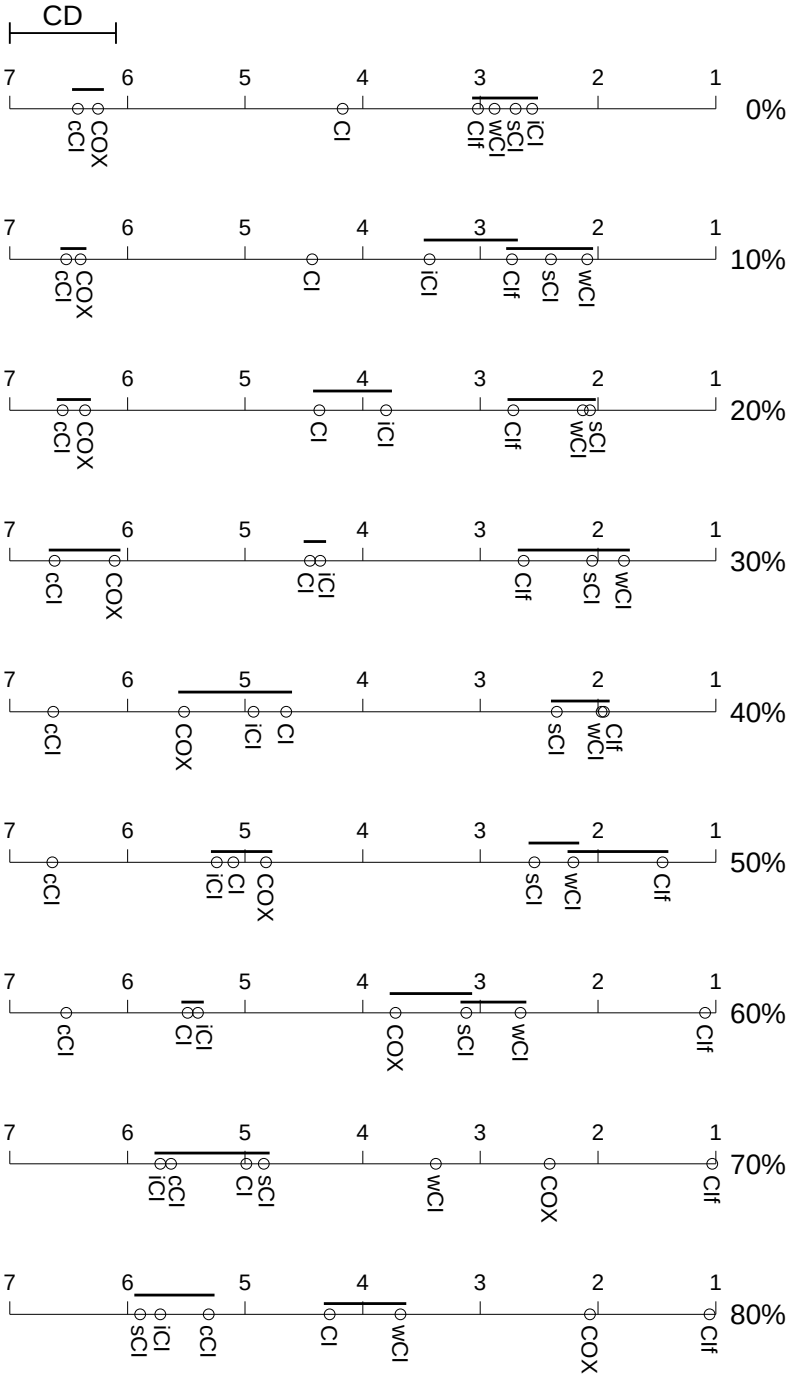
Dodatak C

Rezultati vrednovanja simulacijskom studijom

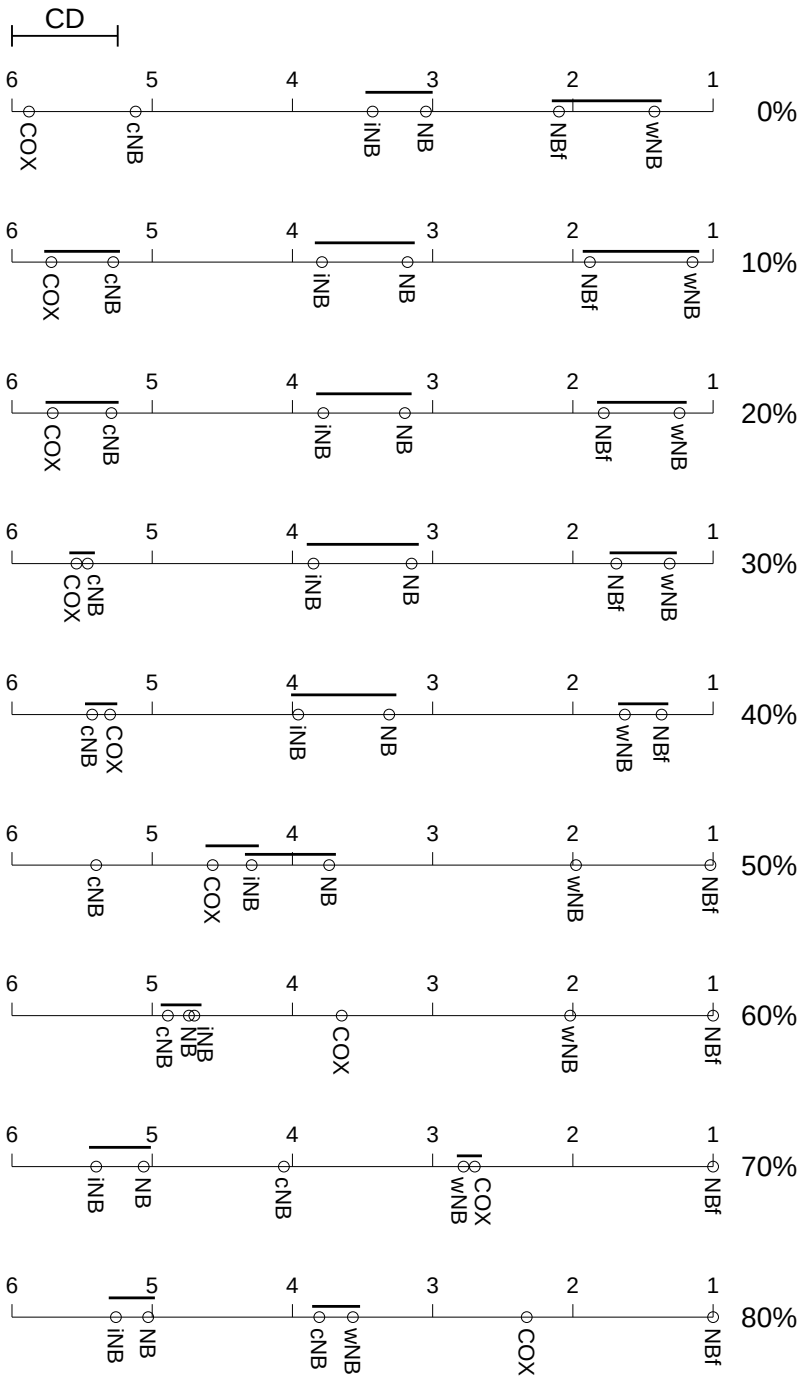
U nastavku su predstavljeni grafovi srednjih rangova rezultata simulacijske studije na ostalim metrikama.



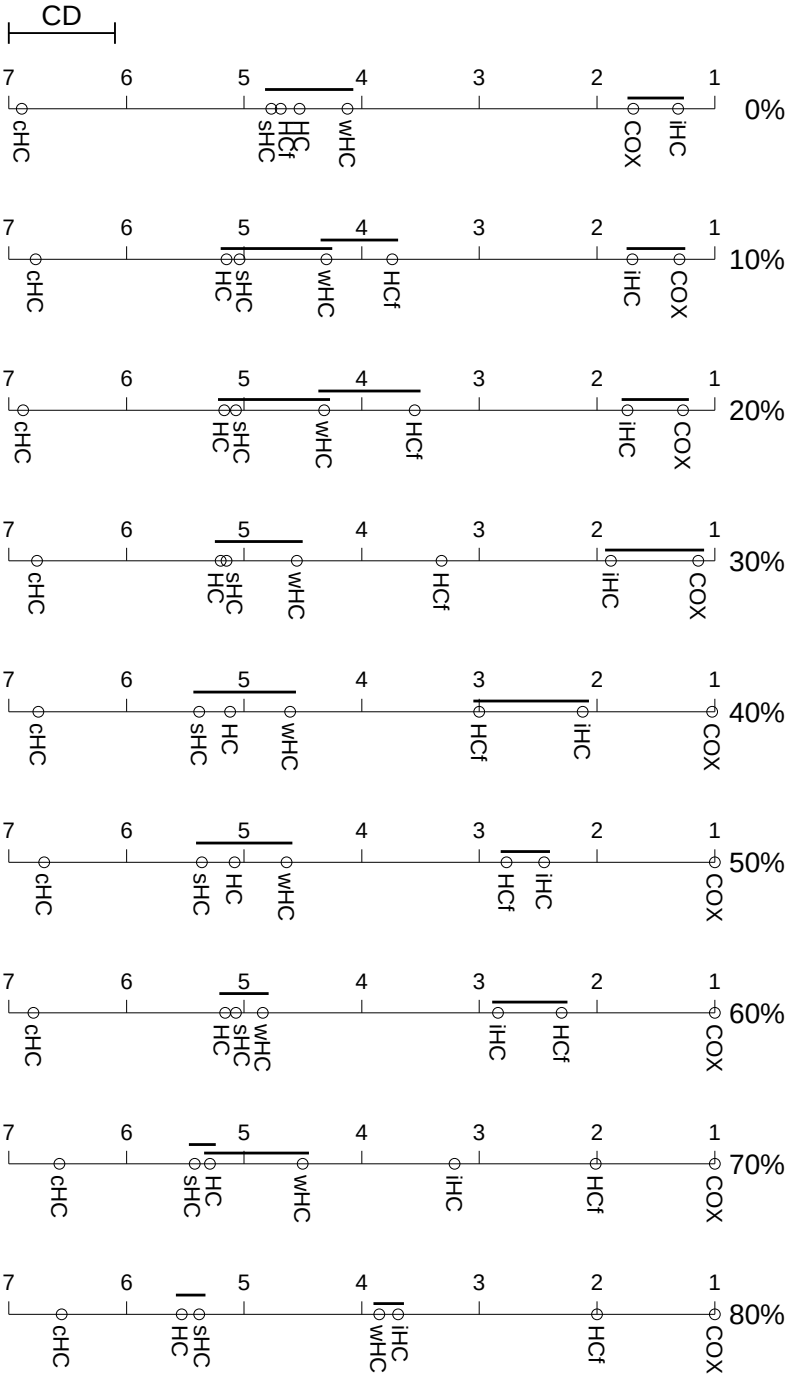
Slika C.1: Srednji rangovi težinske točnosti klasifikacije postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



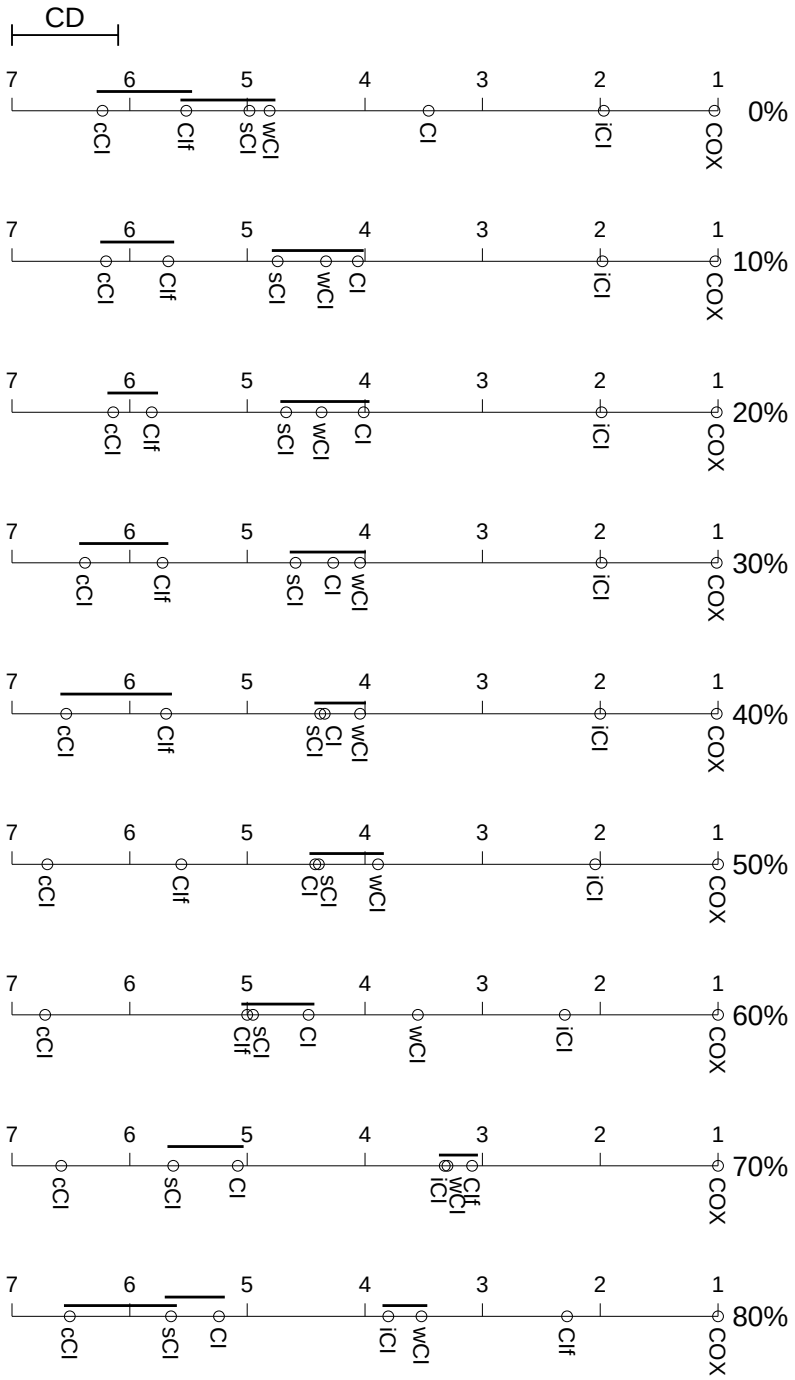
Slika C.2: Srednji rangovi težinske točnosti klasifikacije postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



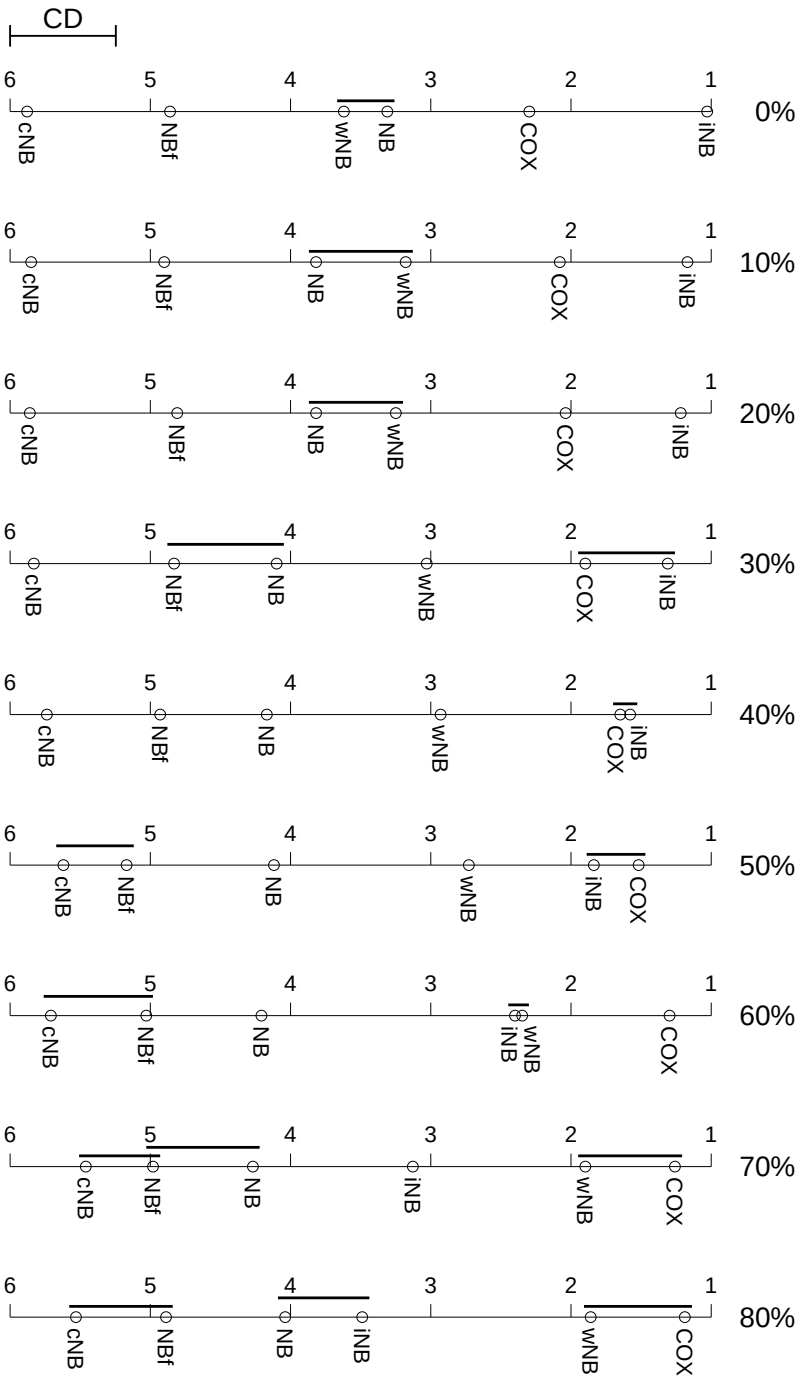
Slika C.3: Srednji rangovi težinske točnosti klasifikacije postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



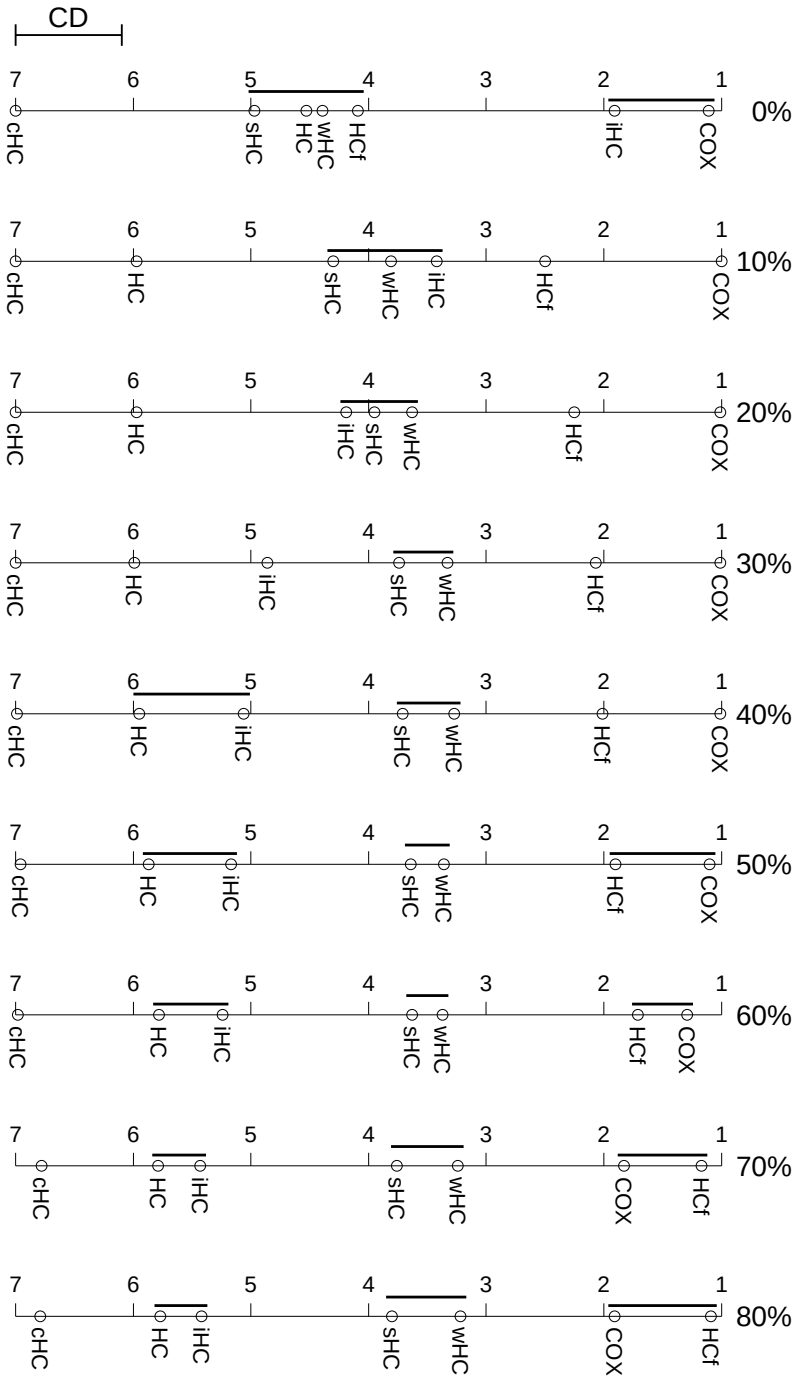
Slika C.4: Srednji rangovi indeksa suglasnosti postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



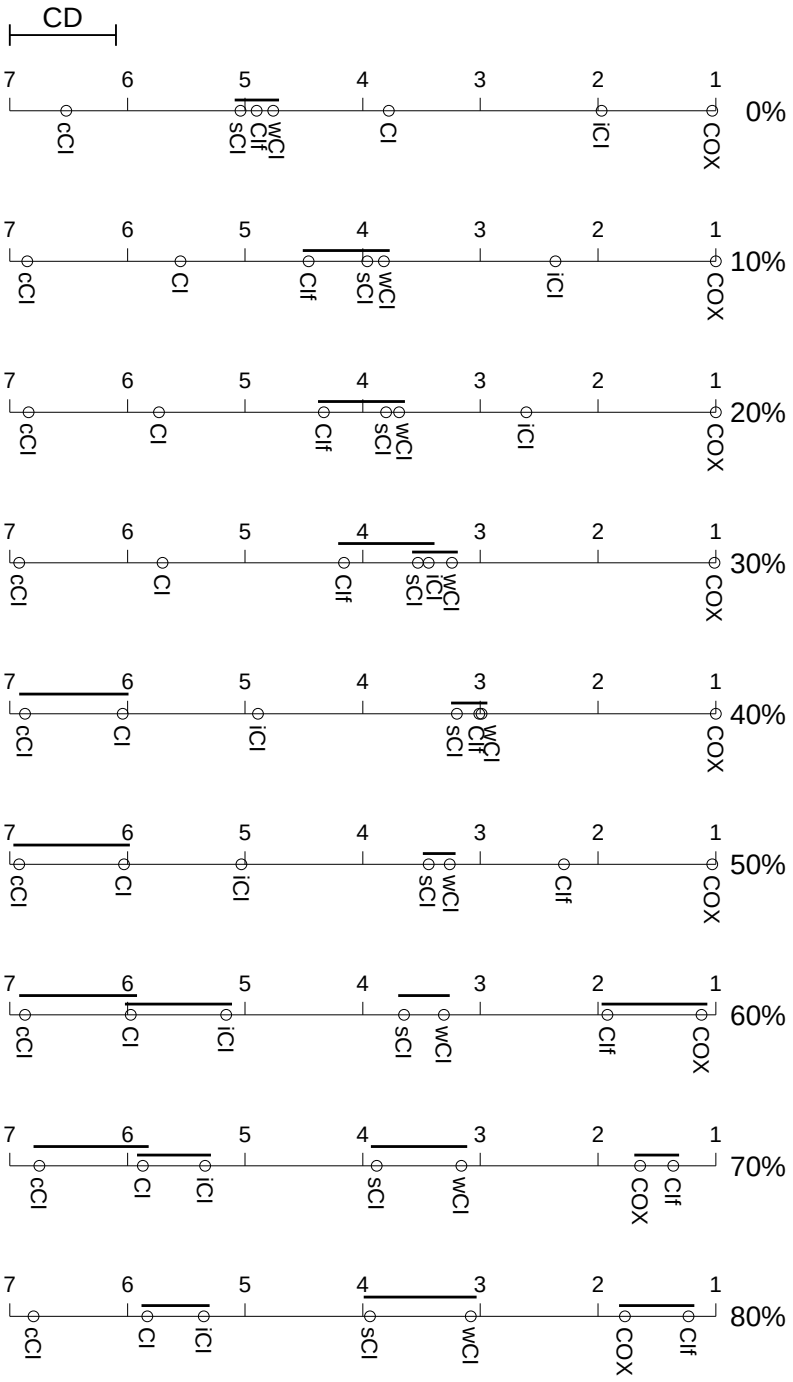
Slika C.5: Srednji rangovi indeksa suglasnosti postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



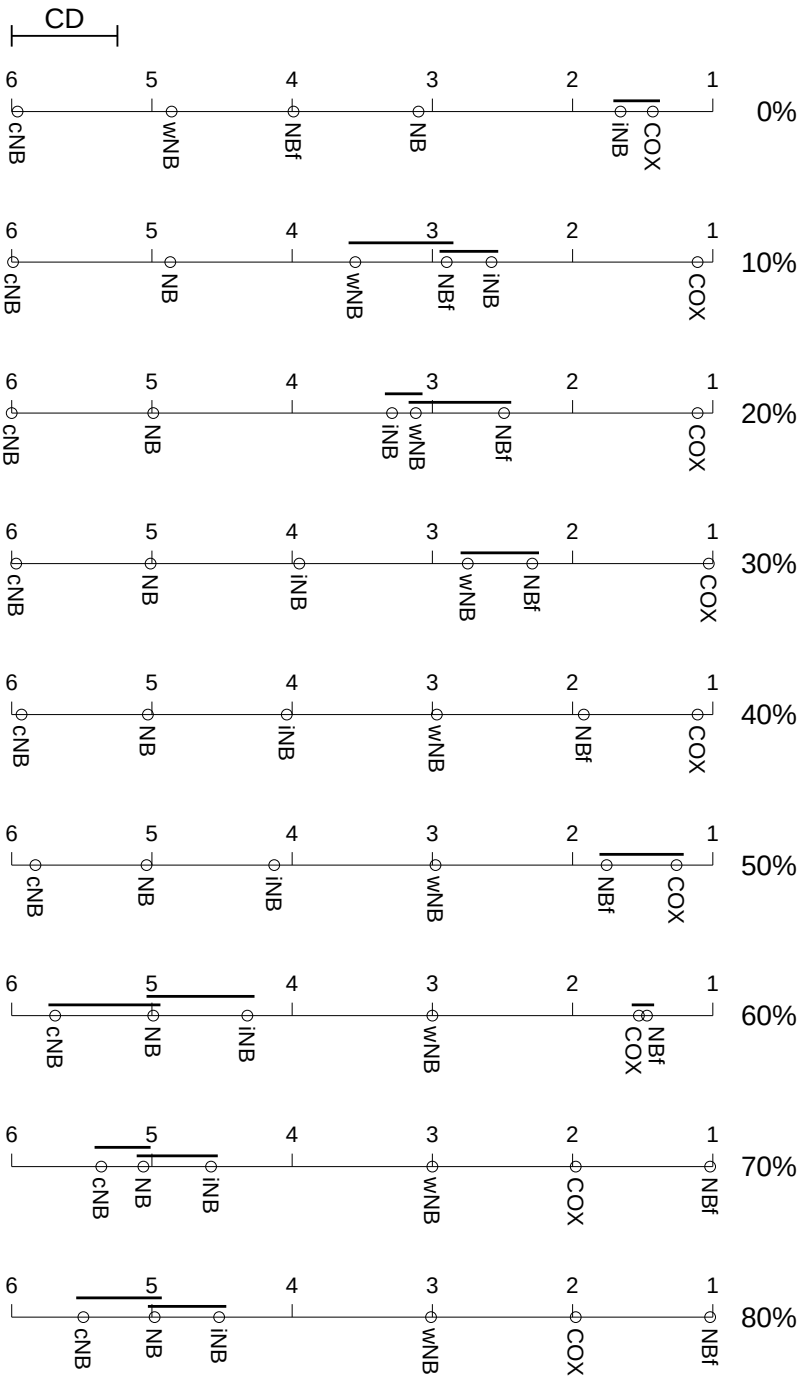
Slika C.6: Srednji rangovi indeksa suglasnosti postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika C.7: Srednji rangovi integrirane Brierove ocjene postupaka učenja Bayesovih mreža algoritmom penjanja uzbrdo i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.



Slika C.8: Srednji rangovi integrirane Brierove ocjene postupaka učenja Bayesovih mreža algoritmom uvjetnih neovisnosti i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.

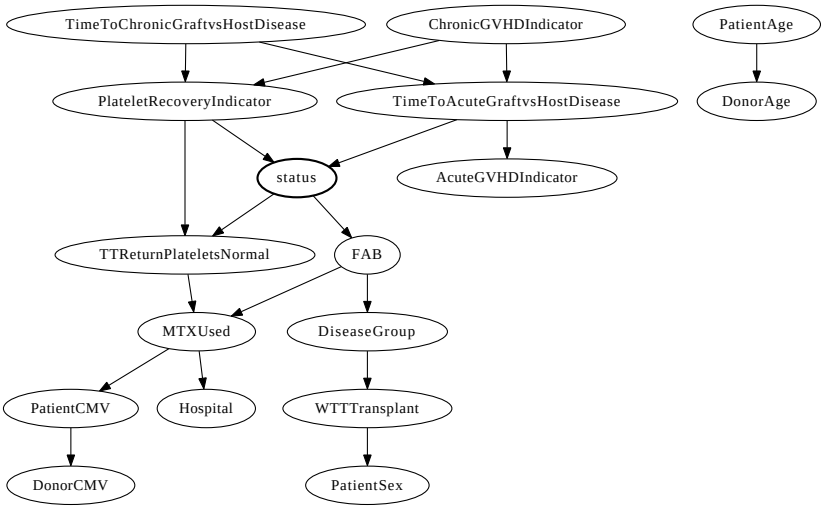


Slika C.9: Srednji rangovi integrirane Brierove ocjene postupaka učenja naivnog Bayesovog klasifikatora i modela proporcionalnih hazarda, za svaku razinu cenzure (postotak s desne strane) posebno. Grupe postupaka koji nisu značajno različiti ($\alpha = 0.05$), to jest onih čiji su srednji rangovi međusobno bliži od kritične razlike CD, spojene su crtama.

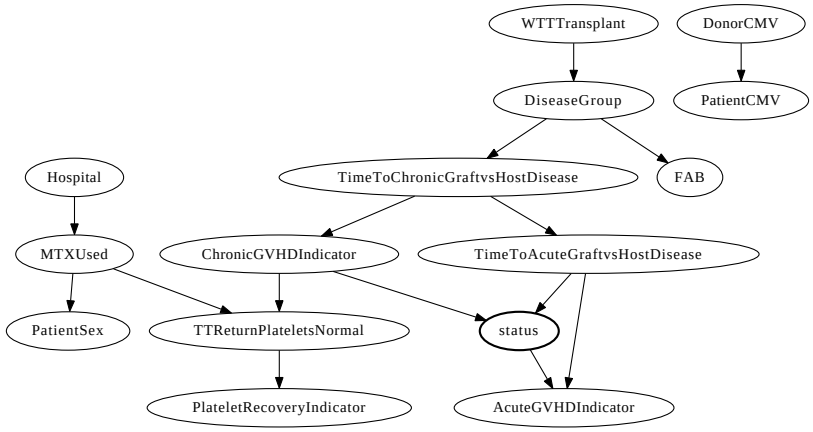
Dodatak D

Bayesove mreže naučene iz realnih domena

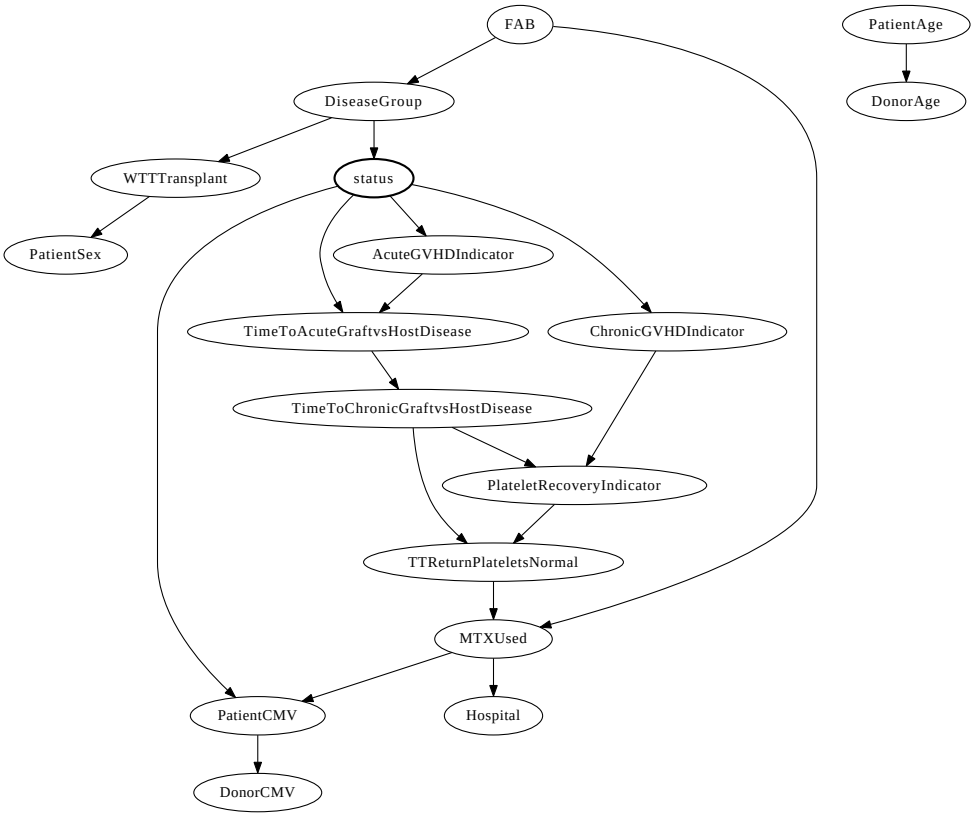
U nastavku su predstavljene Bayesove mreže dobivene različitim postupcima učenja za sve realne domene.



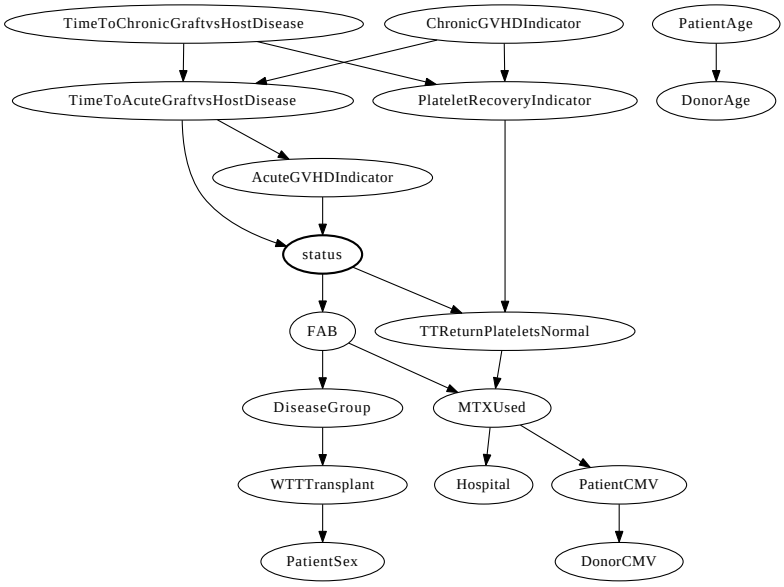
(a) HC i sHC



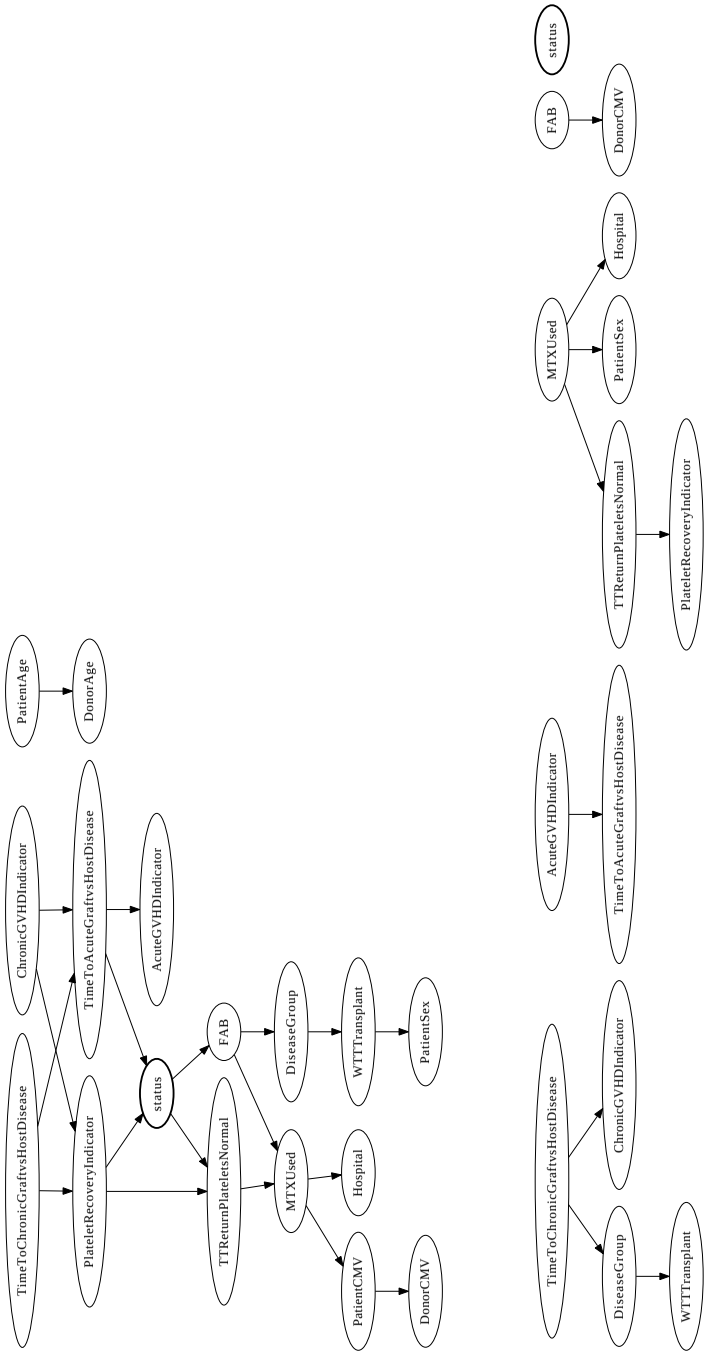
(b) cHC


FAB
PatientAge
DiseaseGroup
DonorAge
WTTTransplant
status
PatientSex
AcuteGVHDIndicator
TimeToAcuteGraftvsHostDisease
ChronicGVHDIndicator
TimeToChronicGraftvsHostDisease
PlateletRecoveryIndicator
TTReturnPlateletsNormal
MTXUsed
PatientCMV
Hospital
DonorCMV

(c) HCf

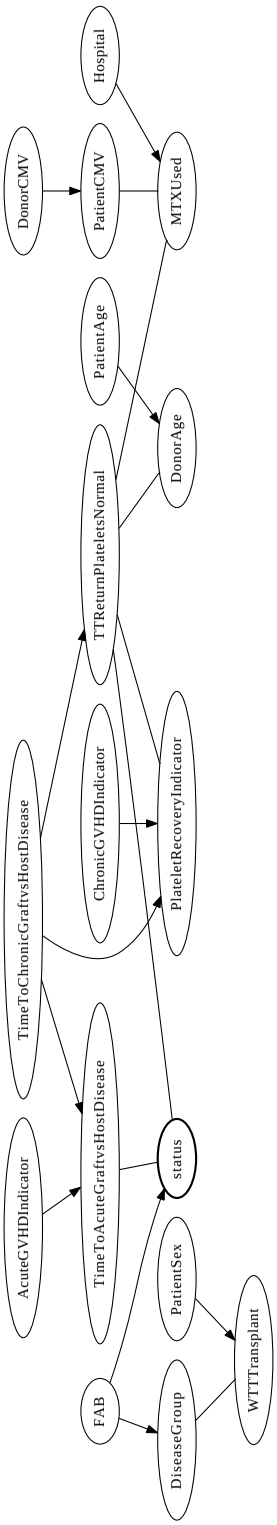

TimeToChronicGraftvsHostDisease
ChronicGVHDIndicator
PatientAge
TimeToAcuteGraftvsHostDisease
PlateletRecoveryIndicator
DonorAge
AcuteGVHDIndicator
status
FAB
TTReturnPlateletsNormal
DiseaseGroup
MTXUsed
WTTTransplant
Hospital
PatientCMV
PatientSex
DonorCMV

(d) wHC

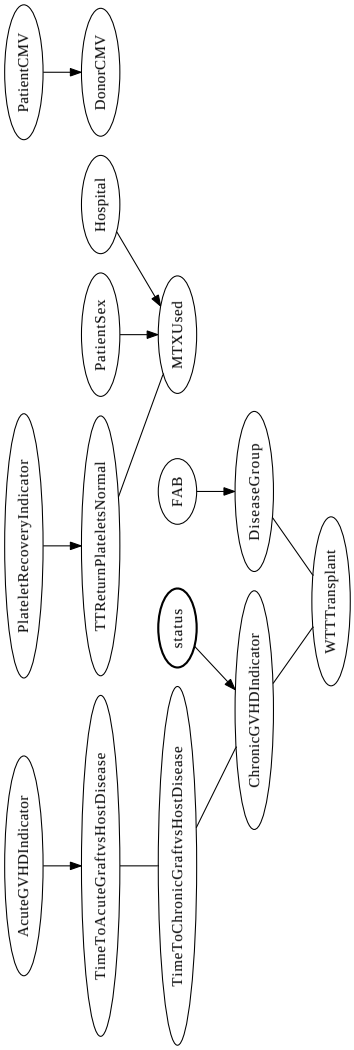


(e) iHC-I (f) iHC-II

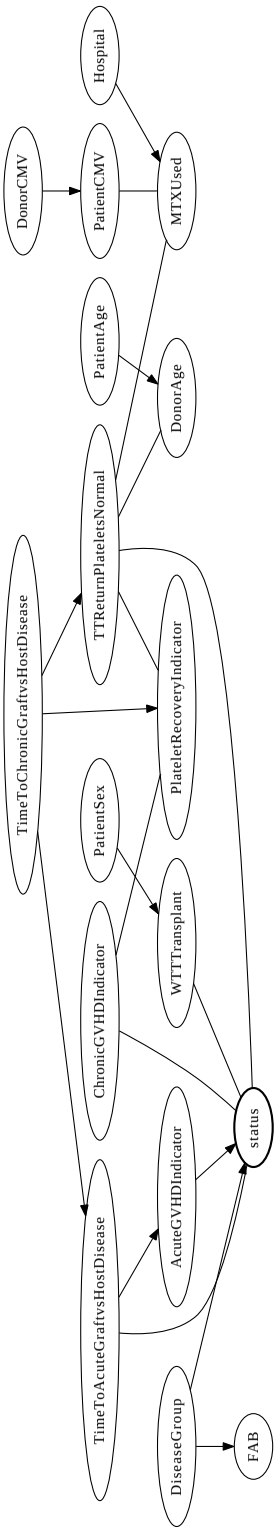
Slika D.1: Bayesove mreže naučene algoritmom penjanja uzbrdo na domeni transplantacije koštane srži.



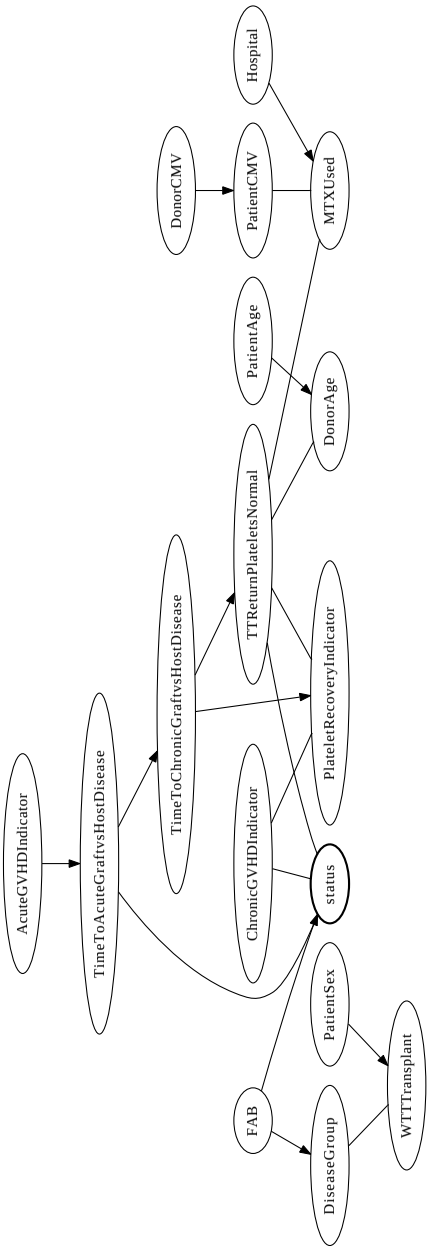
(a) CI i sCI



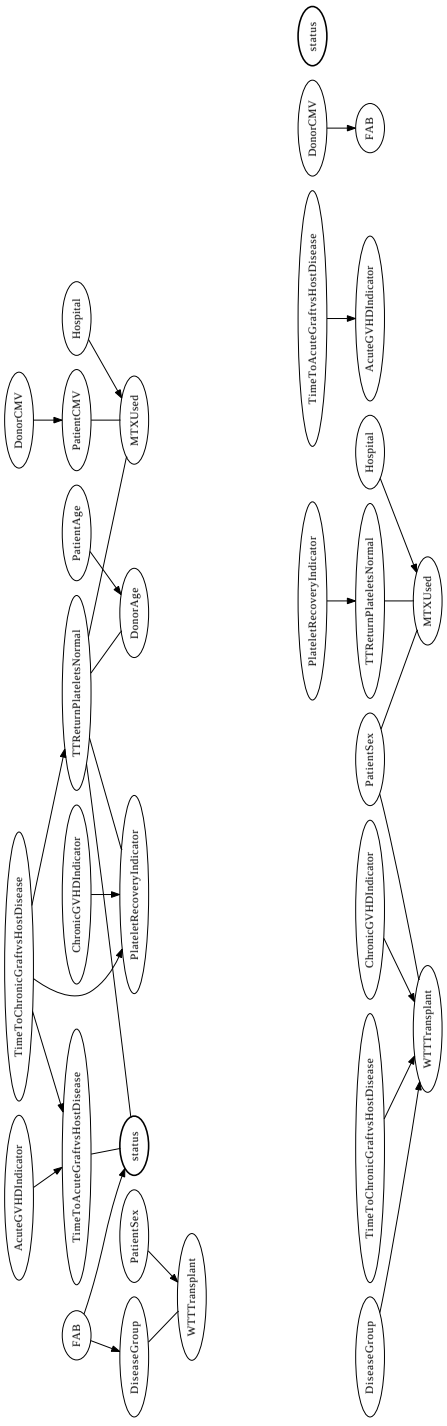
(b) cCI


Hospital
DonorCMV
PatientCMV
MTXUsed
PatientAge
DonorAge
TTReturnPlateletsNormal
TimeToChronicGraftvsHostDisease
PlateletRecoveryIndicator
PatientSex
ChronicGVHDIndicator
WTTTransplant
status
AcuteGVHDIndicator
TimeToAcuteGraftvsHostDisease
DiseaseGroup
FAB

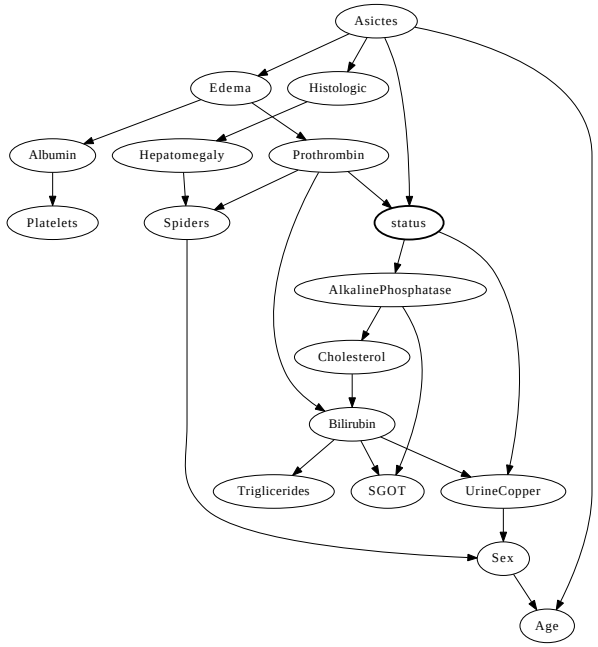
(c) CIf


Hospital
DonorCMV
PatientCMV
MTXUsed
PatientAge
DonorAge
TTReturnPlateletsNormal
TimeToChronicGraftvsHostDisease
PlateletRecoveryIndicator
ChronicGVHDIndicator
status
AcuteGVHDIndicator
TimeToAcuteGraftvsHostDisease
PatientSex
FAB
DiseaseGroup
WTTTransplant

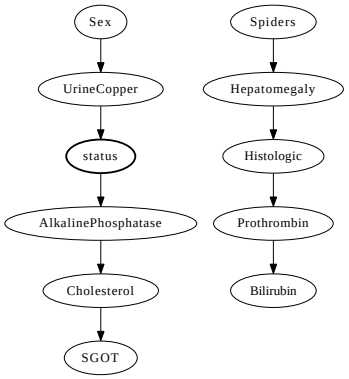
(d) wCI



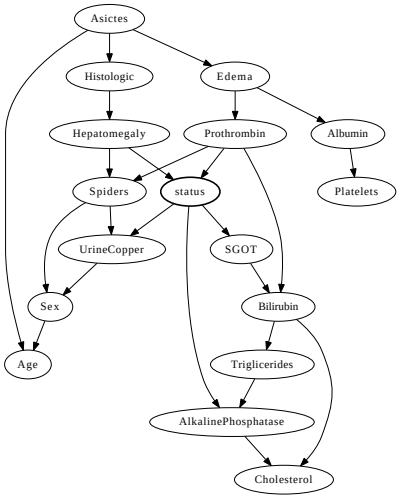
Slika D.2: Bayesove mreže naučene algoritmom uvjetnih neovisnosti na domeni transplantacije koštane srži.



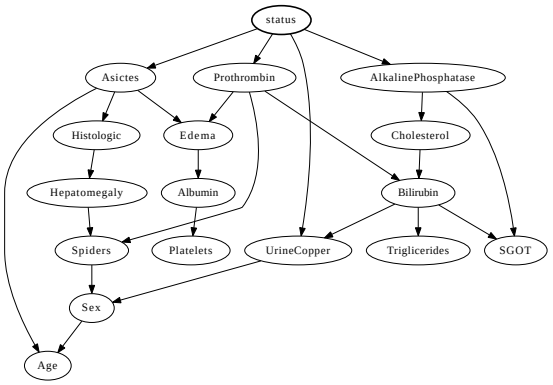
(a) HC i sHC



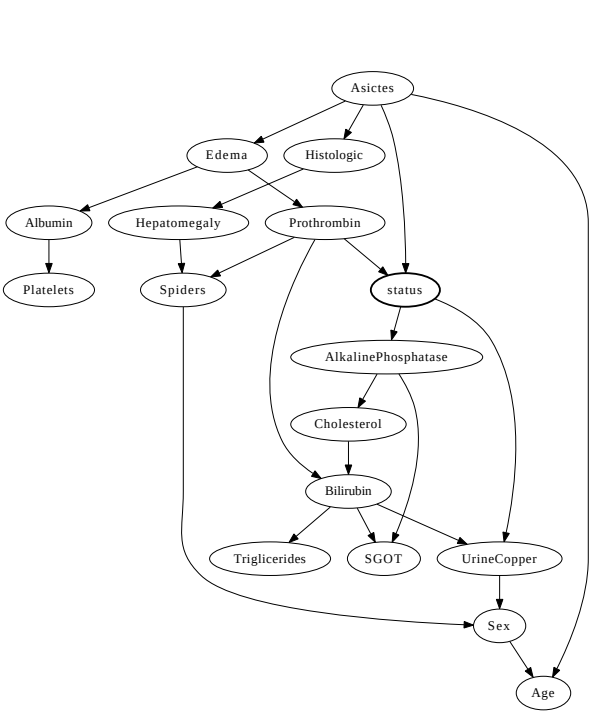
(b) cHC



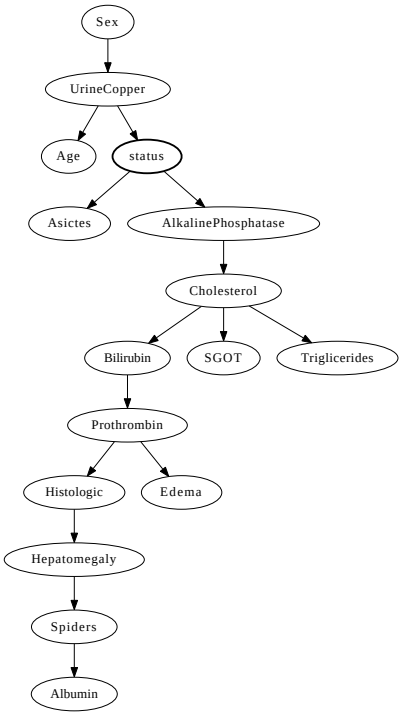
(c) HCf



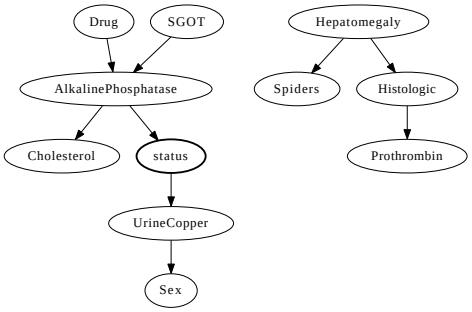
(d) wHC



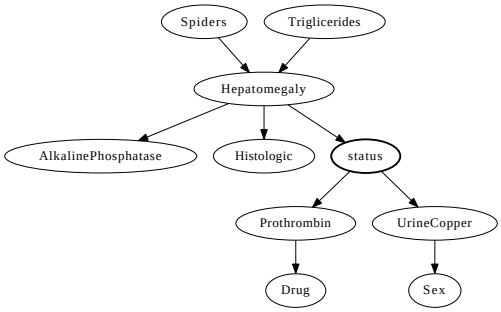
(e) iHC-I



(f) iHC-II

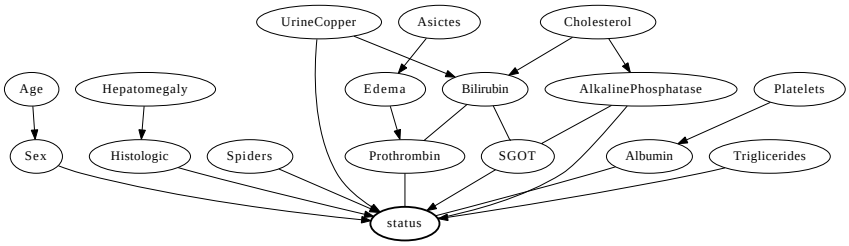


(g) iHC-III

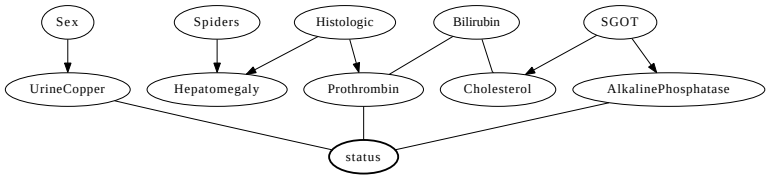


(h) iHC-IIII

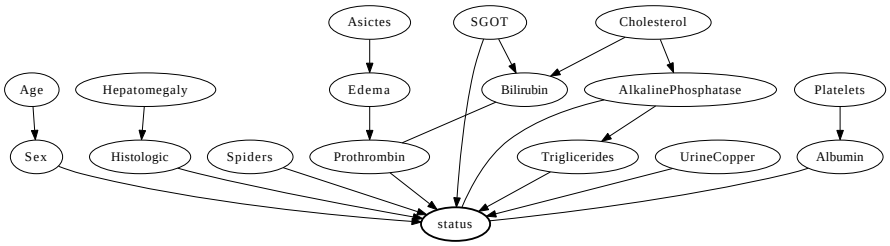
Slika D.3: Bayesove mreže naučene algoritmom penjanja uzbrdo na domeni ciroze jetre.



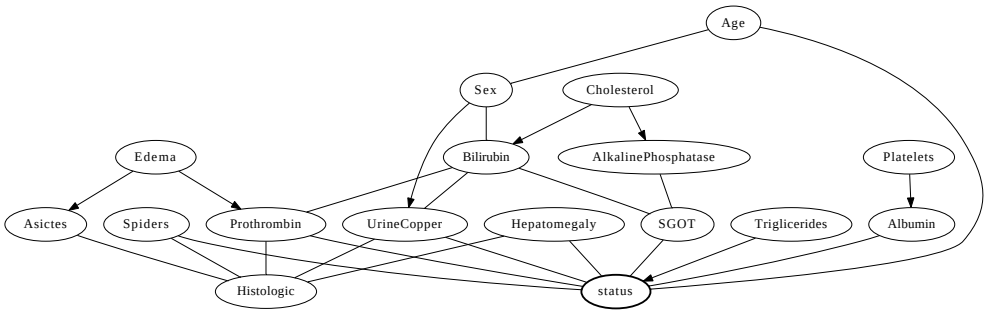
(a) CI i sCI



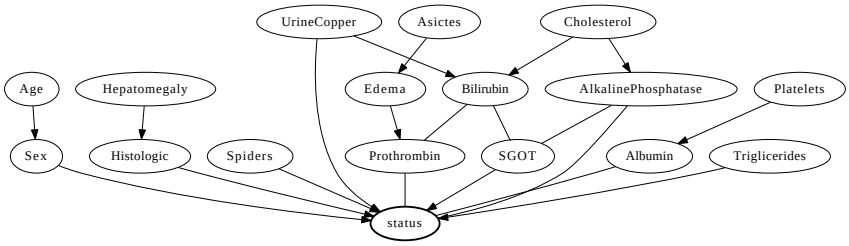
(b) cCI



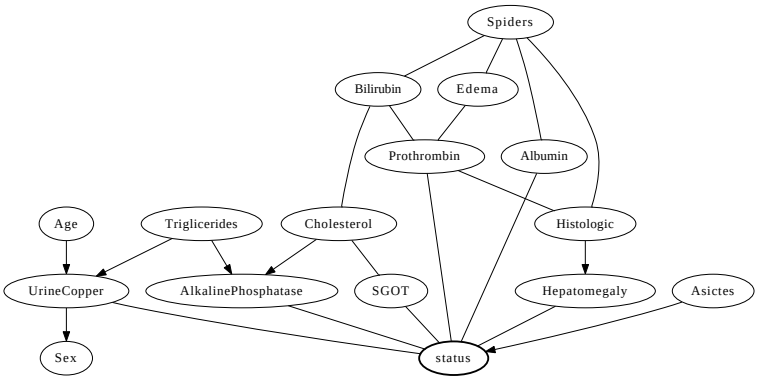
(c) CIf



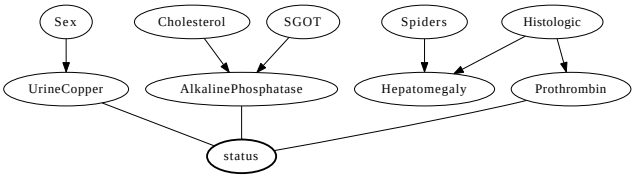
(d) wCI



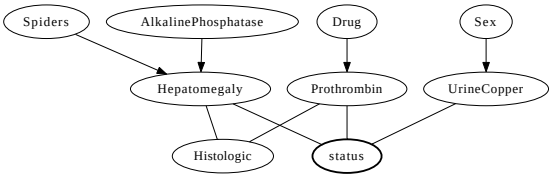
(e) iCI-I



(f) iCI-II

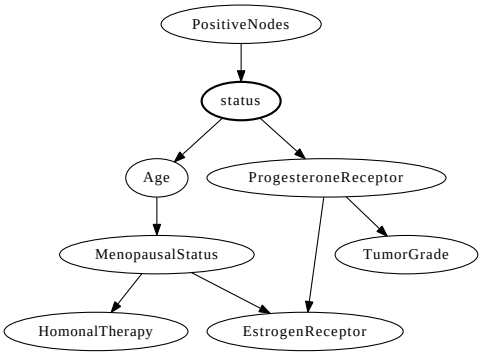


(g) iCI-III

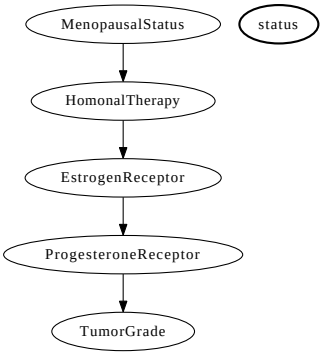


(h) iCI-IIII

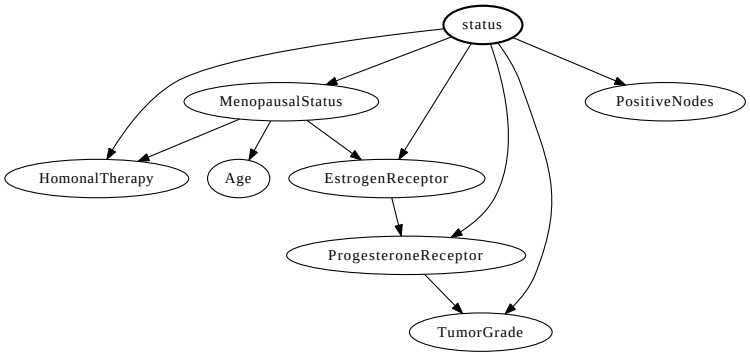
Slika D.4: Bayesove mreže naučene algoritmom uvjetnih neovisnosti na domeni ciroze jetre.



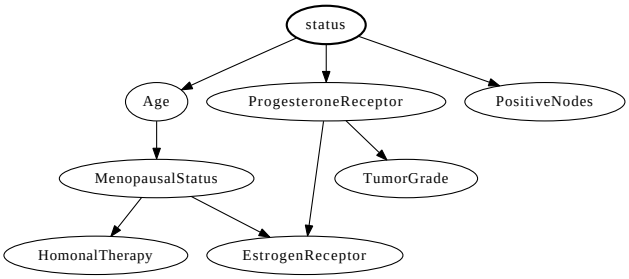
(a) HC i sHC



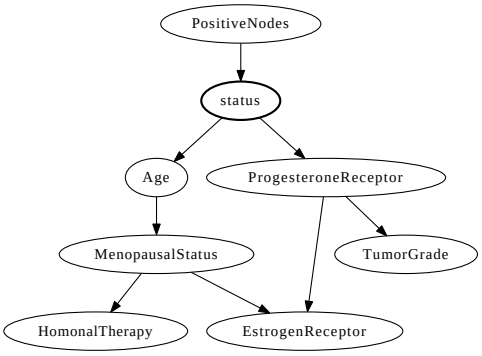
(b) cHC



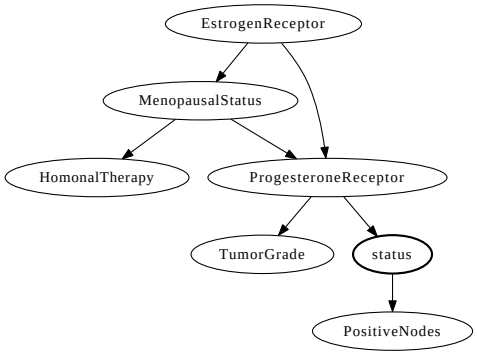
(c) HCf



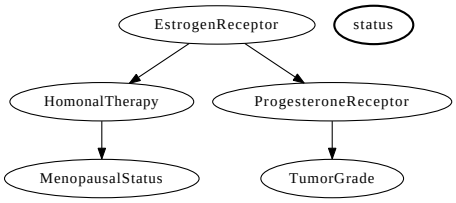
(d) wHC



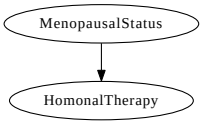
(e) iHC-I

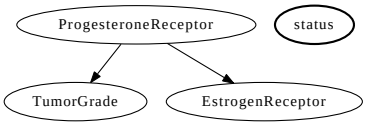


(f) iHC-II

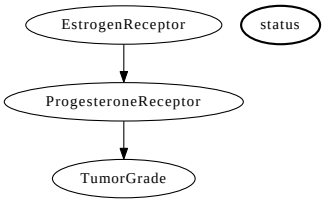


(g) iHC-III



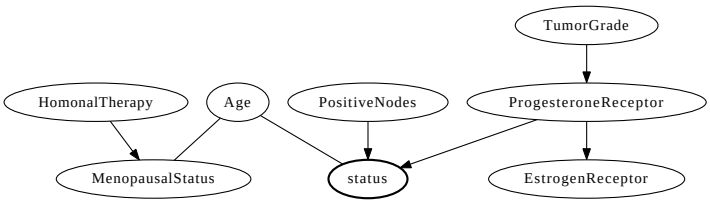


(h) iHC-IIII

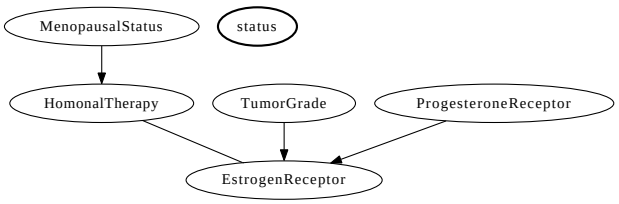


(i) iHC-IIIII

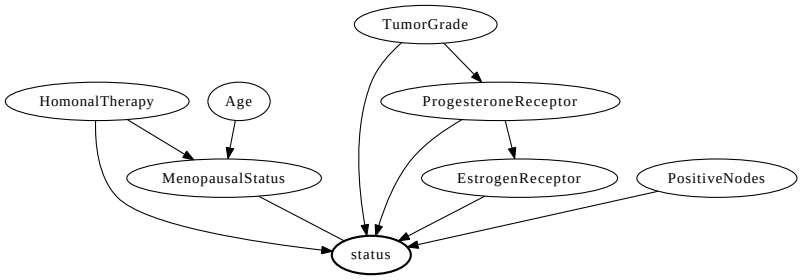
Slika D.5: Bayesove mreže naučene algoritmom penjanja uzbrdo na domeni raka dojke.



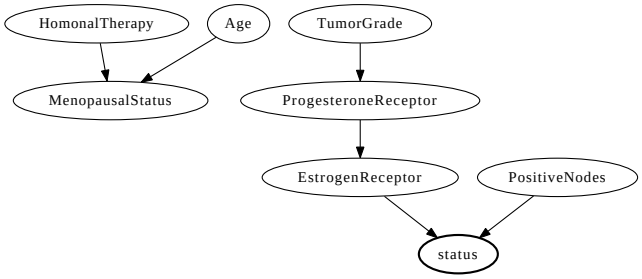
(a) CI i sCI



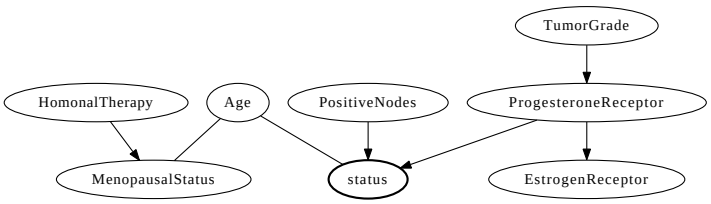
(b) cCI



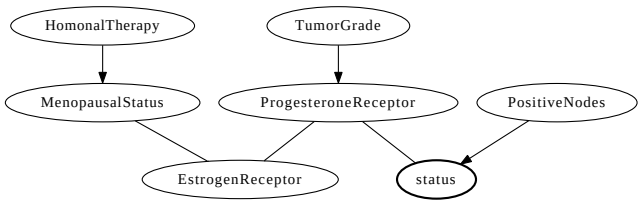
(c) CIf



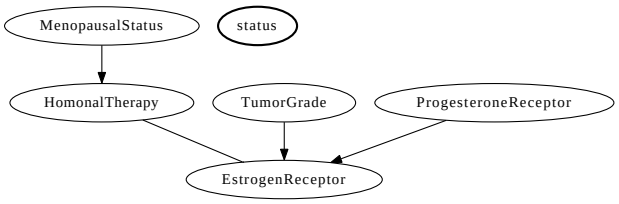
(d) wCI



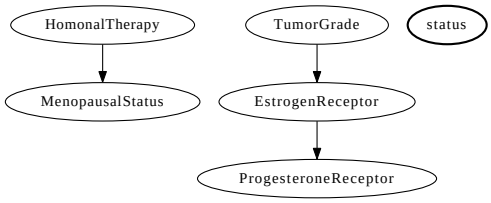
(e) iCI-I



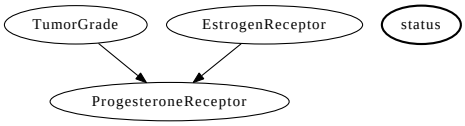
(f) iCI-II



(g) iCI-III



(h) iCI-IIII



(i) iCI-IIIII

Slika D.6: Bayesove mreže naučene algoritmom uvjetnih neovisnosti na domeni raka dojke.

Popis literature

Abellan, J., Gomez-Olmedo, M. & Moral, S. (2006). Some Variations on the PC Algorithm. In M. Studený & J. Vomlel, eds., *Proceedings of the Third European Workshop on Probabilistic Graphical Models*, Reprostredisko UK MFF, Prague, Czech Republic. 39

Alpaydin, E. (1999). Combined 5×2 cv F Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, **11**, 1885–1892. 50

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge, MA, USA. 121

Andreassen, S., Riekehr, C., Kristensen, B., Schønheyder, H.C. & Leibovici, L. (1999). Using Probabilistic and Decision–Theoretic Methods in Treatment and Prognosis Modeling. *Artificial Intelligence In Medicine*, **15**, 121–134. 2

Bender, R., Augustin, T. & Blettner, M. (2005). Generating Survival Times to Simulate Cox Proportional Hazards Models. *Statistics in Medicine*, **24**, 1713–1723. 55

Berthold, M. & Hand, D.J. (2003). *Intelligent Data Analysis: an Introduction*. Springer, New York, NY, USA. 121

Biganzoli, E., Boracchi, P., Mariani, L. & Marubini, E. (1998). Feed Forward Neural Networks for the Analysis of Censored Survival Data: a Partial Logistic Regression Approach. *Statistics in Medicine*, **17**, 1169–1186. 21, 28

Bishop, C.M. (2007). *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA. 16, 38, 121

Blanco, R., Inza, I., Merino, M., Quiroga, J. & Larrañaga, P. (2005). Feature Selection in Bayesian Classifiers for the Prognosis of Survival of Cirrhotic Patients Treated with TIPS. *Journal of Biomedical Informatics*, **38**, 376–388. 22

Borgelt, C. & Kruse, R. (2002). *Graphical Models: Methods for Data Analysis and Mining*. John Wiley & Sons, Chichester, United Kingdom. 37, 39, 121

Braunwald, E., Zipes, D.P. & Libby, P. (2001). *Heart Disease: a Textbook of Cardiovascular Medicine*. W.B. Saunders Company, Philadelphia, PA, USA. 5

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, USA. 1, 16

Brier, G.W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1–3. 47

Bromberg, F. & Margaritis, D. (2009). Improving the Reliability of Causal Discovery from Small Data Sets Using Argumentation. *Journal of Machine Learning Research*, **10**, 301–340. 39

Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, D.E., Weinstein, J.N., Harrell Jr., F.E., Marks, J.R., Winchester, D.P. & Bostwick, D.G. (1997). Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction. *Cancer*, **79**. 21, 27

Cheng, J., Greiner, R., Kelly, J., Bell, D. & Liu, W. (2002). Learning Bayesian Networks from Data: an Information–Theory Based Approach. *Artificial Intelligence*, **137**, 43–90. 39

Chickering, D.M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, **3**, 507–554. 39

Clarke, J. & West, M. (2008). Bayesian Weibull Tree Models for Survival Analysis of Clinico–Genomic Data. *Statistical Methodology*, **5**, 238–262. 1

Contal, C. & O'Quigley, J. (1999). An Application of Changepoint Methods in Studying the Effect of Age on Survival in Breast Cancer. *Computational Statistics and Data Analysis*, **30**, 253–270. 54, 88

Cooper, G.F. & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, **9**, 309–347. 2, 38, 39

Cox, D.R. (1972). Regression Models and Life–Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220. 1, 14, 44

Delen, D., Walker, G. & Kadam, A. (2005). Predicting Breast Cancer Survivability: a Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine*, **34**, 113–127. 21, 27, 30

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1–30. 49, 50

Domingos, P. & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero–One Loss. *Machine Learning*, **29**, 103–130. 43

Duda, R.O., Hart, P.E. & Stork, D.G. (2001). *Pattern Classification*. John Wiley & Sons, Hoboken, NJ, USA, 2nd edn. 1

Ellson, J., Gansner, E., Koutsofios, L., North, S. & Woodhull, G. (2002). Graphviz – Open Source Graph Drawing Tools. *Lecture Notes in Computer Science*, 483–484. 3

Evers, L. & Messow, C.M. (2008). Sparse Kernel Methods for High–Dimensional Survival Data. *Bioinformatics*, **24**, 1632–1638. 22

Fayyad, U.M. & Irani, K.B. (1993). Multi–Interval Discretization of Continuous-Valued Attributes for Classification Learning. In R. Bajcsy, ed., *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022–1027, Morgan Kaufmann, San Francisco, CA, USA. 87

Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Hoboken, NJ, USA. 95

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 675–701. 49

Friedman, N. (1998). The Bayesian Structural EM Algorithm. In G.F. Cooper & S. Moral, eds., *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 129–138, Morgan Kaufmann, San Fransisco, CA, USA. 2

Friedman, N. & Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, **50**, 95–125. 39

Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Statistics in Medicine*, **18**, 2529–2545. 47

Grünwald, P.D., Myung, I.J. & Pitt, M.A. (2005). *Advances in Minimum Description Length: Theory and Applications*. The MIT Press. 41

Hamming, R.W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, **29**, 147–160. 22

Hand, D. & Yu, K. (2001). Idiot's Bayes: Not So Stupid after All? *International Statistical Review*, **69**, 385–398. 43

Hanley, J.A. & McNeil, B.J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29–36. 46

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. & Rosati, R.A. (1982). Evaluating the Yield of Medical Tests. *Journal of the American Medical Association*, **247**, 2543–2546. 46

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA. 54, 70

Heckerman, D. (1996). A Tutorial on Learning With Bayesian Networks. Tech. rep., Microsoft Research. 38

Heckerman, D., Geiger, D. & Chickering, D.M. (1995). Learning Bayesian Networks: the Combination of Knowledge and Statistical Data. *Machine Learning*, **20**, 197–243. 2, 39, 40

Hoot, N. & Aronsky, D. (2005). Using Bayesian Networks to Predict Survival of Liver Transplant Patients. In C.P. Friedman, J. Ash & P. Tarczy-Hornoch, eds., *AMIA Annual Symposium Proceedings*, vol. 2005, 345–349, American Medical Informatics Association, Bethesda, MD, USA. 2

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A. & van der Laan, M.J. (2006). Survival Ensembles. *Biostatistics*, **7**, 355–373. 22

Hruschka, E. & Ebecken, N. (2007). Towards Efficient Variables Ordering for Bayesian Networks Classifier. *Data & Knowledge Engineering*, **63**, 258–269. 39

Iman, R.L. & Davenport, J.M. (1980). Approximations of the Critical Region of the Friedman Statistic. *Communications in Statistics - Theory and Methods*, **9**, 571–595. 50

Janžura, M. & Nielsen, J. (2006). A Simulated Annealing–Based Method for Learning Bayesian Networks from Statistical Data. *International Journal of Intelligent Systems*, **21**, 335. 41

Jerez-Aragonés, J., Gómez-Ruiz, J., Ramos-Jiménez, G., Muñoz-Pérez, J. & Alba-Conejo, E. (2003). A Combined Neural Network and Decision Trees Model for Prognosis of Breast Cancer Relapse. *Artificial Intelligence In Medicine*, **27**, 45–63. 28

Kaplan, E.L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of American Statistical Association*, **53**, 457–481. 12, 30, 46

Kattan, M.W., Hess, K.R. & Beck, J.R. (1998). Experiments to Determine Whether Recursive Partitioning (CART) or an Artificial Neural Network Overcomes Theoretical Limitations of Cox Proportional Hazards Regression. *Computers and Biomedical Research*, **31**, 363–373. 26, 27

Kjaerulff, U.B. & Madsen, A.L. (2007). *Bayesian Networks and Influence Diagrams: a Guide to Construction and Analysis*. Springer, New York, NY, USA. 7, 10

Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, NY, USA, 2nd edn. 10, 54, 88, 89

Kleinbaum, D.G. (2005). *Survival Analysis: a Self–Learning Text*. Springer, New York, NY, USA, 2nd edn. 1, 18

Kooperberg, C., Stone, C.J. & Truong, Y.K. (1995). Hazard Regression. *Journal of the American Statistical Association*, **90**, 78–94. 1

Krause, P.J. (1998). Learning Probabilistic Networks. *The Knowledge Engineering Review*, **13**, 321–351. 36

Kronek, L.P. & Reddy, A. (2008). Logical Analysis of Survival Data: Prognostic Survival Models by Detecting High–Degree Interactions in Right–Censored Data. *Bioinformatics*, **24**, 248–253. 22

Lam, W. & Bacchus, F. (1994). Learning Bayesian Belief Networks: an Approach Based on the MDL Principle. *Computational Intelligence*, **10**, 269–293. 2, 39, 41

Lee, E.T. & Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Hoboken, NJ, USA, 3rd edn. 1, 11, 12, 15, 16, 44

Lisboa, P.J.G., Wong, H., Harris, P. & Swindell, R. (2003). A Bayesian Neural Network Approach for Modelling Censored Data with an Application to Prognosis after Surgery for Breast Cancer. *Artificial Intelligence In Medicine*, **28**, 1–25. 28

Lucas, P. & Abu-Hanna, A. (1999). Prognostic Methods in Medicine. *Artificial Intelligence in Medicine*, **15**, 105–119. 1

Lucas, P., Boot, H. & Taal, B. (1998). Computer–Based Decision Support in the Management of Primary Gastric non–Hodgkin Lymphoma. *Methods of Information in Medicine*, **37**, 206–219. 2

Lucas, P.J.F., de Bruijn, N.C., Schurink, K. & Hoepelman, A. (2000). A Probabilistic and Decision–Theoretic Approach to the Management of Infectious Disease at the ICU. *Artificial Intelligence In Medicine*, **19**, 251–279. 2

Lucas, P.J.F., van der Gaag, L.C. & Abu-Hanna, A. (2004). Bayesian Networks in Biomedicine and Health–Care. *Artificial Intelligence In Medicine*, **30**, 201–214. 2, 22

Marshall, A., McClean, S., Shapcott, M. & Millard, P. (2000). Learning Dynamic Bayesian Belief Networks Using Conditional Phase–Type Distributions. *Lecture Notes in Computer Science*, 516–523. 22

Michie, D., Spiegelhalter, D.J., Taylor, C.C. & Campbell, J. (1995). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA. 48

Mozina, M., Demšar, J., Kattan, M. & Zupan, B. (2004). Nomograms for Visualization of Naive Bayesian Classifier. *Lecture Notes in Computer Science*, 337–348. 43

Murphy, K.P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California. 22

Nemenyi, P. (1963). *Distribution-free Multiple Comparisons*. Ph.D. thesis, Princeton University. 50

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, USA. 1, 2, 7, 10, 35, 37, 38, 40

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK. 2, 39

Peña-Reyes, C.A. & Sipper, M. (2000). Evolutionary Computation in Medicine: an Overview. *Artificial Intelligence In Medicine*, **19**, 1–23. 22

Peto, R. & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society*, **135**, 185–207. 15

Polikar, R. (2006). Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, **6**, 21–45. 28

Quinlan, J.R. (1986). Induction of Decision Trees. *Machine Learning*, **1**, 81–106. 16

R Development Core Team (2008). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT, `http://www.R-project.org` (dostupno: 21. prosinca 2009). 3, 102

Ripley, B.D. & Ripley, R.M. (2001). Neural Networks as Statistical Methods in Survival Analysis. In V. Gant & R. Dybowski, eds., *Clinical Applications of Artificial Neural Networks*, 237–255, Cambridge University Press, Cambridge, UK. 21, 27, 31, 46

Royston, P. & Sauerbrei, W. (2004). A New Measure of Prognostic Separation in Survival Data. *Statistics in Medicine*, **23**, 723–748. 25, 53, 55, 70

Russell, S.J. & Norvig, P. (2002). *Artificial Intelligence: a Modern Approach*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edn. 41

Schumacher, M., Bastert, G., Bojar, H., Hubner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R.L. & Rauschecker, H.F. (1994). Randomized 2×2 Trial Evaluating Hormonal Treatment and the Duration of Chemotherapy in Node–positive Breast Cancer Patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, **12**, 2086–2093. 102

Sheskin, D. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC press, Boca Raton, FL , USA. 50

Sierra, B. & Larranaga, P. (1998). Predicting Survival in Malignant Skin Melanoma Using Bayesian Networks Automatically Induced by Genetic Algorithms. An Empirical Comparison Between Different Approaches. *Artificial Intelligence in Medicine*, **14**, 215–230. 22

Snow, P.B., Smith, D.S. & Catalona, W.J. (1994). Artificial Neural Networks in the Diagnosis and Prognosis of Prostate Cancer: a Pilot Study. *The Journal of urology*, **152**, 1923–1926. 21, 26

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA. 39

Štajduhar, I. & Bratko, I. (2007). Likelihood Based Classification in Bayesian Networks. In V. Devedžic, ed., *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, 335–340, ACTA Press, Calgary, Canada. 31

Štajduhar, I., Dalbelo-Bašić, B. & Bogunović, N. (2009). Impact of Censoring on Learning Bayesian Networks in Survival Modelling. *Artificial Intelligence in Medicine*, **47**, 199–217. 25, 26

Tsamardinos, I., Brown, L. & Aliferis, C. (2006). The Max–Min Hill–Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, **65**, 31–78. 39

Verma, T. & Pearl, J. (1992). An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation. In D. Dubois & M.P. Wellman, eds., *Proceedings of the 8th Annual Conference on Uncertainty in Artificial Intelligence*, 323–330, Morgan Kaufmann, San Fransisco, CA, USA. 40

Witten, I.H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, USA. 3, 48

Xie, X. & Geng, Z. (2008). A Recursive Method for Structural Learning of Directed Acyclic Graphs. *Journal of Machine Learning Research*, **9**, 459–483. 39

Zupan, B., Demšar, J., Kattan, M.W., Beck, R. & Bratko, I. (2000). Machine Learning for Survival Analysis: a Case Study on Recurrence of Prostate Cancer. *Artificial Intelligence in Medicine*, **20**, 59–75. 21, 28, 30

Popis slika

Popis tablica

Popis simbola

Analiza preživljenja

t	Vrijeme
T	Vrijeme praćenja (preživljenja)
$S(t)$	Funkcija preživljenja
$\mathscr{X}_i$	Rizična skupina

Bayesova mreža

$\mathcal{B}$	Bayesova mreža
$\mathcal{G}$	Usmjereni aciklički graf
$\mathbf{A}$	Skup lukova
$\mathbf{V}$	Skup čvorišta
A_{ij}	Luk koji povezuje čvorište V_i s čvorištem V_j
V_i	Čvorište
$\boldsymbol{\pi}_i$	Skup roditelja čvorišta V_i
n	Broj čvorišta
r_i	Kardinalnost čvorišta V_i
q_i	Produkt kardinalnosti čvorišta skupa $\boldsymbol{\pi}_i$

Podaci

$\mathcal{D}$	Podaci, skup primjeraka
$\mathbf{X}$	Skup kovarijata
X_i	Kovarijata
O	Varijata od interesa, razred, status

$\mathbf{x}_i$ Primjerak

m Broj kovarijata

N Broj primjeraka u uzorku

Ostalo

$\mathcal{M}$ Model

P Vjerojatnost

$\mathbf{P}$ Skup vjerojatnosti

$\mathcal{L}$ Funkcija izglednosti

$\rho_{i,j}$ Korelacijski koeficijent

β_i Regresijski koeficijent

α Razina statističke značajnosti testa

p Empirijska razina statističke značajnosti

Naslov: UČENJE BAYESOVIH MREŽA IZ CENZURIRANIH PODATAKA O PREŽIVLJENJU

Sažetak:
Bayesove mreže kao jedan od oblika predstavljanja znanja i nadziranog strojnog učenja, stekle su veliku popularnost zbog učinkovitog zaključivanja i mogućnosti intuitivnog grafičkog predstavljanja uzročno-posljedičnih veza među kovarijatama. Zbog tih karakteristika često se koriste u sustavima potpore odlučivanju u raznim poljima medicine, no njihova je primjena gotovo u potpunosti zanemarena u analizi preživljenja. Razlog tome leži u potrebi za korištenjem posebnih postupaka rukovanja podacima o preživljenju zbog lošeg utjecaja cenzure. Ova se disertacija bavi problemom učenja Bayesovih mreža iz cenzuriranih podataka o preživljenju. Predstavljeno je više poznatih i jedan novi postupak pripremne obrade takvih podataka za potrebe strojnog učenja. Korišteni modeli su dva poznata algoritma za učenje Bayesovih mreža, model naivnog Bayesovog klasifikatora i regresijski model proporcionalnih hazarda. Provedeno je temeljito testiranje simulacijskom studijom i na nekoliko realnih domena iz područja kliničke medicine. Procjena učinkovitosti postupaka pripremne obrade podataka na modelima izvedena je statističkom usporedbom rezultata testova više standardnih metrika strojnog učenja i metrika analize preživljenja. Pored toga, testirana je i sposobnost ispravnog otkrivanja uzročne strukture algoritama za učenje Bayesovih mreža uz različite postupke pripremne obrade podataka. Rezultati simulacijske studije sugeriraju kako je predložen postupak pripremne obrade podataka odstranjivanjem šuma cenzure, izvrsno rješenje za visokocenzurirane domene.

Ključne riječi: Bayesova mreža, analiza preživljenja, cenzura, strojno učenje, prognostički model, predstavljanje znanja.

Title: LEARNING BAYESIAN NETWORKS FROM CENSORED SURVIVAL DATA

Abstract:
As a form of knowledge representation and supervised machine learning, Bayesian networks have become increasingly popular because of their efficient inference and their inherent ability of intuitive graphical representation of causal interactions among covariates. For these reasons they are often used in decision support systems in various fields of medicine. At the same time they are being almost completely ignored in survival analysis. This is because special data-handling techniques are needed for handling survival data, due to the bad influence of censoring. This thesis deals with the problem of learning Bayesian networks from censored survival data. Several known data-preprocessing techniques for machine learning and a new technique are presented. Models used include two well-known algorithms for learning Bayesian networks, the naive Bayes classifier and the proportional hazards regression model. Thorough testing was conducted on both a simulation study and on several real-world domains from the field of clinical medicine. Assessment of the efficiency of the data-preprocessing techniques on models was conducted using statistical comparison on test results of several standard machine-learning and survival-analysis metrics. The Bayesian network learning algorithms' ability of identifying the correct causal structure after using different data-preprocessing techniques was also tested. Simulation study results suggest that the proposed data-preprocessing technique of reducing censoring noise performs excellently in heavily censored domains.

Keywords: Bayesian network, survival analysis, censoring, machine learning, prognostic model, knowledge representation.

Životopis

Ivan Štajduhar rođen je 26. svibnja 1978. godine u Rijeci. Gimnaziju je završio u Rijeci, prirodoslovno-matematičko usmjerenje. Godine 1996. upisuje se na Fakultet računarstva i informatike (Fakulteta za računalništvo in informatiko) u Ljubljani, gdje diplomira u rujnu 2001. godine na smjeru informatika. Poslijediplomski studij za stjecanje akademskog stupnja magistra znanosti upisuje 2001. godine na istom Fakultetu, na smjeru informacijski sustavi i odlučivanje, pod mentorstvom akad. prof. dr. Ivana Bratka. Uspješno je obranio magistarski rad "Učenje Bayesovih mreža iz medicinskih podataka" 12. prosinca 2005. Skoro dvije godine radio je kao izvođač informacijskih rješenja u poduzeću ComLand d.o.o. u Ljubljani, uglavnom na razvoju *e-commerce* aplikacija i internet stranica za državnu upravu. U kolovozu 2002. stupa u radni odnos s Tehničkim fakultetom u Rijeci u svojstvu znanstvenog novaka. Istovremeno je radio na održavanju dijela informacijskog sustava poduzeća Marand d.o.o, od 2002. do 2005. Kao istraživač radio je na projektima MZOŠ 0069015 "Raspodijeljeni sustavi upravljanja proizvodnim procesima korištenjem agenata" i MZOŠ 069-0362214-1575 "Optimizacija i dizajn vremensko-frekvencijskih distribucija".

Resume

Ivan Štajduhar was born in Rijeka on May 26 1978. He graduated from a local high school, a gymnasium of mathematics, physics and computer science. In 1996 he enrolled the University of Ljubljana Faculty of computer and information science, where he studied information science until graduation, September 2001. Immediately after graduation, in order to earn the title of Master of Science, he enrolled the Postgraduate study of Information and decision systems at the same university under the mentorship of Professor Ivan Bratko, PhD. He successfully defended his thesis "Learning Bayesian networks from medical data" on December 12 2005. For nearly two years he provided information solutions for ComLand Ltd. Ljubljana, mainly developing e-commerce application software and web sites for the state administration. In August 2002 he started working at the Faculty of engineering in Rijeka as a junior scientist. Simultaneously he worked as an information systems administrator for a company called Marand Ltd., from 2002 until 2005. As a researcher he worked on MZOŠ project no. 0069015 "Distributed control systems for managing manufacturing processes using agents" and project no. 069-0362214-1575 "Optimization and design of time-frequency distributions".