Skriveni Markov modeli u medicini primenjeni u analizi vremenskih nizova podataka za otkrivanje upale pluća pneumonije u bolnicama

UDK: 007:616-083 ; 519.863 ; 519.246.8

Marek Opuszko[1], Johannes Ruhland[1], Franziska Oroszi[2], Michael Hartmann[2] · Martin Specht[2] [1]Friedrich-Schiller-
Univerzitet, Jena (FSU), Nemačka, Odsek za informacione sisteme
marek.opuszko@uni-jena.de, j.ruhland@wiwi.uni-jena.de
[2]Univerzitetska bolnica, Jena, Nemačka

XII Internacionalni Simpozijum SymOrg 2010, 09.-12. Jun 2010, Zlatibor, Srbija

Pneumonija – kao zapaljenjsko oboljenje pluća – predstavlja opasnu i često fatalnu bolest. Od ventilatorne pneumonije (VAP), posebnog oblika ove bolesti, oboljeva otprilike jedna petina pacijenata u odeljenjima intenzivne nege (ICU). Na osnovu skupa podataka za period od dve godine, prikupljenih u jednom velikom odeljenju intenzivne nege, ispitivali smo novi metod obrade vremenskih serija podataka da bismo razvili sistem koji bi na vreme upozorio na opasnost od oboljevanja od pneumonije. U ovom sistemu fokusiramo se na period pre početka bolesti i pokušavamo da odredimo budući tok događaja. Prilikom kategorisanja i predviđanja datih vremenskih serija kod pacijenta koristili smo Skriveni Markov Model (HMM) i paradigmu slaganja. Na kraju smo prikazali primer sa skupom podataka za stvarnog pacijenta i na taj način pokazali kakva je korist od našeg pristupa.

1. Uvod

Sve veća primena informacionih tehnologija u bolnicama i drugim medicinskim ustanovama imala je za posledicu potrebu da se već uskladišteni podaci istraže kako bi se više saznalo o oboljenjima, toku bolesti, mogućim metodama lečenja, itd. Posebno kritične sredine kao što su odeljenja intenzivne nege veoma su zainteresovane za ovu temu. U ovim odeljenjima 8-20% pacijenata oboli od VAP (N.N., 2005) i to vodi stopi smrtnosti od 20-50% ili čak 70% (N.N., 2005) (Heyland i dr., 1999) (Tejerina i dr., 2006). Stoga je veoma značajno da se postavi rana i precizna dijagnoza VAP. Precizna dijagnoza – a samim tim i brže ozdravljanje – skraćuje boravak pacijenta u ICU i smanjuje kako nepotreban stres pacijentu tako i troškove bolnice koji bi se mogli izbeći (Oroszci, 2008). Pored toga, lekari su pretrpani ogromnim brojem podataka koje moraju da zapisuju svakodnevno. Stoga su potrebni novi metodi rada. Metod izdvajanja podataka nudi mogućnost da se čisti i sirovi podaci pretvore u znanje što bi doprinelo većoj uspešnosti metoda lečenja.

Ovaj rad je deo interdisciplinarnog projekta u kome učestvuju istraživači sa odseka informacionih sistema Univerziteta Friedrich-Schiller (FSU) u Jeni, odeljenje za intenzivnu negu (ICU) i bolnička apoteka iste ove institucije (Oroszci, 2008). Ovaj projekat u kom je primenjen standardni proces obrade u tehnici izdvajanja podataka (CRISP-DM) (Chapman i dr., 2000) ima za cilj da primeni tehnike izdvajanja podataka na datoteku ICU. U ovom radu posebno ćemo se baviti obradom vremenskih serija podataka, što je značajan istraživački obuhvat za ceo projekat. Podaci koje analiziramo su vremenske serije podataka agregatne vrednosti koje su nastale tokom rane faze projekta (Oroszci, 2008). Poseban cilj ovog rada i drugih istraživanja u okviru ovog projekta bio je da se identifikuju pogodni dodaci za predviđanje pneumonije koji će se koristiti u daljim istraživanjima. Suštinsko pitanje koje se postavlja u vezi sa ovim projektom glasi: Da li postoje razlike u pred-fazi bolesti između pacijenata koji boluju od pneumonije i onih koji ne boluju od ove bolesti? Sledeće pitanje koje se postavlja jeste da, ako razlike postoje, da li su one trivijalne, na primer „ako merena vrednost dostigne određenu tačku, pneumonija će se manifestovati narednog dana“ ili da li tok bolesti sadrži i složenije obrasce koje treba otkriti? Stoga, ako ovi obrasci postoje, metode izdvajanja podataka će možda moći da ih iskoriste za formiranje jednog sistema za rano upozoravanje. Kad uspostavlja dijagnozu, ordinirajući lekar suočava se sa različitim vrstama podataka i inputima informacija. Kad su mu sve dostupniji kompjuterski obrađeni i digitalno sačuvani podaci, potrebni su mu i alati kojima će efektivno i efikasno da obradi ovaj input. Sistem za rano upozoravanje mora da ponudi pouzdane i razumljive informacije koji će biti od pomoći lekarima u svakodnevnom poslu, da bi postavili najbolju dijagnozu primenom sistema dijagnostičke podrške.

Struktura ovog rada je sledeća:
U poglavlju 2 predstavljamo pregled datih podataka i njihovu strukturu. U poglavlju 3 ukratko ilustrujemo teoriju i funkcionalnost Skrivenin Markov Modela. Simulacijom u poglavlju 4 predstavljamo sposobnost sistema da prikaže boravak pacijenta u ICU. Sledi poglavlje 5 u kome predstavljamo probnu varijantu kojom pokazujemo kako komponente međusobno delu-

ju. Na kraju rada predstavljamo rezultate naših istraživanja i naglašavamo neke mogućnosti daljih istraživanja.

2. Podaci

Svi podaci prikupljeni su u početnoj fazi projekta godine 2004. i 2005. i već su prošli kroz preliminarnu obradu. Ceo skup podataka obuhvatio je više od 4000 varijabli. Nažalost, ne postoji ni jedna jedinstvena klinička manifestacija na osnovu koje bi se dijadnostifikovao VAP, ali smo primenili nekoliko metoda čiji su se rezultati razlikovali (Rea-Neto i dr., 2008). Koncentrisali smo se na zbir kliničkih pulmonarnih infekcija (CPIS) koji je izračunat za svakog pacijenta tokom 2004. i 2005. godine. CPIS je zbirna vrednost koja je formirana da bi se olakšalo dijagnostifikovanje pneumonije, a prvi ju je predložio Pugin i dr., 1991. godne. Iako CPIS ima određena ograničenja koja se vezuju za njegovu umerenu uspešnost, on predstavlja dobar alat za dijagnostifikovanje VAP (Rea-Neto i dr., 2008).

Tabela 1: *Vrednovanje inputa CPIS*

Ulazna odlika	Zbir		
	0	1	2
Sekrecija iz traheja	Retka	Obilna	Gnojna
Radiografski infiltrati	Nema	Neujednačeni ili difuzni	Lokalizovani
Visoka temperatura(°C)	≤ 36,5 i ≥ 38,4	> 38,4 i ≥ 38,9	> 38,9 ili < 36
Leukocitoza	≤ 4.000 i ≥ 11.000	< 4.000 ili > 11.000	(> 4.000 ili < 11.000) i ≤ 500 trakasti oblici
Oksidacija (PaO2/FIO2)	> 240 ili precizan sindrom respiratornog distresa (ARDS)		≥ 240 i bez ARDS
Mikrobiologija	Negativna		Pozitivna

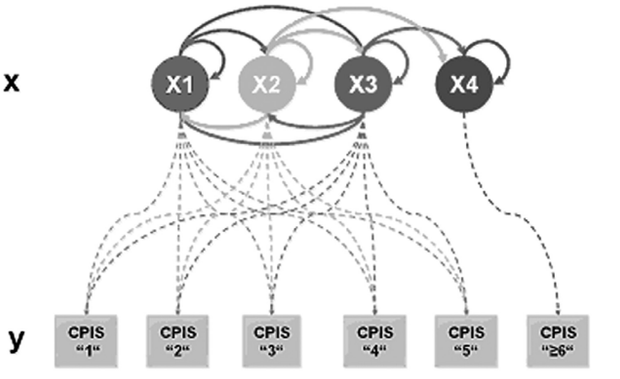
Kako je prikazano na tabeli 1, CPIS predstavlja skup izražen celim brojem, koji sadrži 6 komponenti (sekrecija iz traheja, radiografski infiltrati, visoka temperatura, oksidacija i polu-kvantitativne kulture aspirate/udahnute tvari, mikrobiologija) (Pugin i dr., 1991). Svaka komponenta dodaje vrednost celog broja između 0 i 2. Stoga CPIS ima maksimalnu vrednost 12 – ako sve odlike imaju vrednost 2 – a minimalnu vrednost 0. Prema međunarodnoj praksi, smatramo da je dijagnoza pneumonije konstatovana ako CPIS dostigne vrednost ≥ 6 (Rea-Neto i dr., 2008). Prvi dan pneumonije naziva se „dan reakcije". Na osnovu ove konvencije mogu se identifikovati dve grupe slučajeva, slučajevi sa pneumonijom i slučajevi koji nemaju pneumoniju. Ovu informaciju iskoristićemo kasnije, kad budemo vrednovali naš model. Prvi pregled podataka pokazao je nepovoljnu distribuciju, posebno u grupi koja boluje od pneumonije. U ovoj grupi mogli smo da izdvojimo 325 vremenskih serija CPIS za godine 2004. i 2005. Zahvaljujući činjenici da se većina merenih vrednosti pojavila u okviru perioda pošto je dan reakcije već dostignut, u ovoj grupi mogli smo da uzmemo u obzir za obradu samo 79 vremenskih serija sa ukupno 425 vrednosti CPIS. Naš cilj je bio da analiziramo fazu pre nastanka pneumonije; podaci za ovaj rani period toka bolesti su od suštinske važnosti. U grupi koja nije imala pneumoniju imali smo dovoljno podataka. Na primer, izdvojili smo 147 slučajeva sa ukupno 995 pojedinačnih vrednosti CPIS za 2004. godinu i 138 slučajeva sa ukupno 827 pojedinačnih vrednosti CPIS za godinu 2005. Pored toga, različite vremenske serije imale su prekide i podaci su bili proređeni zato što su mnogi pacijenti ostajali u bolnici veoma kratko vreme. Šta više, podaci su bili veoma neuravnoteženi pošto je grupa koja nema pneumoniju predstavljena isuviše velikim brojem članova. Povrh svega, period koji je prošao pre nego što se stiglo do dana reakcije bio je veoma kratak u većini slučajeva sa pneumonijom. Zbog ovih ograničenja bili su nam potrebni metodi kojima se može obraditi i ova vrsta podataka.

3. Skriveni Markov modeli

Pristup u istraživanju mogućnosti za predviđanje pneumonije na osnovu vremenskih serija CPIS uglavnom koristi Skriveni Markov model (HMM). Veliki broj drugih metoda teško mogu da obrade vremenske serije arbitrarne i različite dužine, ali HMM nudi mogućnost obrade vremenskih serija čije karakteristike zaista predstavljaju izazov. Modeli HMM decenijama su

uspešno primenjivani u izdvajanju podataka u shvatanju govora (Rabiner, 1989)(Manning and Schütze, 2005) kao i u mnogobrojnim drugim istraživanjima, na primer, u bioinformatici (Gascuel and Moret, 2001) (Bystroff and Krogh, 2008). Matematički opis HMM uglavnom se drži formulacije koju su dali Rabiner i Juang, 1986. Stohastički model HMM karakteriše kombinacija dva nasumična, slučajna procesa. Početni proces sa N različitih stanja $X = \{X1, .., XN\}$ nije vidljiv („skriven je"). Ovaj proces se ne može meriti, ali postoji M emisija $Y = \{y1, .., yM\}$ koje se mogu posmatrati i koje daju informaciju o prvobitnom (originalnom) procesu. Zdravstveno stanje pacijenta može se smatrati nasumičnim, slučajnim procesom. U okviru ovog procesa zdravstveno stanje pacijenta se s vremena na vreme menja. Obično se njegovo zdravstveno stanje može opisati kao „dobro", „loše", „stabilno", isl. Očigledno je stoga da je ovaj proces teško operacionalizovati i neposredno meriti. Svaki lekar koristi simptome i druge informacije do kojih može da dođe da bi postavio dijagnozu o fizičkom stanju pacijenta. Naša pretpostavka sada jeste da se zdravstveno stanje može predstaviti kao nevidljivo (skriveno) stanje HMM.

U tipičnom okruženju HMM se koriste za klasifikovanje vremenskih signala kao što je neprekidni, vezani govor ili genetski niz. Uobičajeno se signal deli na blokove (okvire) u fazi pred-obrade. U našem slučaju, jedna vrednost CPIS predstavlja jedan blok. Zbog ograničenja što je merenje CPIS moguće samo jednom dnevno, okvirna stopa je jedan dan[1]. Stoga celu vremensku seriju CPIS možemo da tumačimo kao emisione simbole koji predstavljaju merljive simptome. Ove vremenske serije omogućavaju nam da vršimo procene u vezi sa – skrivenim – prvobitnim procesom i, pored toga, o zdravstvenom stanju. Na slici 1 prikazana je struktura HMM koja predstavlja tok bolesti. U postavci ove provere pretpostavili smo da postoje četiri originalna stanja koja smo nazvali: „zeleno (X1)" – zdravstveno stanje je stabilno, „žuto (X2) – zdravstveno stanje nije stabilno, „narandžasto (X3) – zdravstveno stanje pacijenta je ozbiljno ugroženo i „crveno" (X4) za stanje oboljenja koje se već manifestuje. Da bi model razvoja bolesti mogao da se napravi, potrebna su nam tri stanja (nije potrebno „crveno"). Slučajna varijabla x(t) predstavlja skriveno stanje u vremenu $t(x(t) \in \{X1, X2, X3, X4\})$. Svako stanje $X1,\ldots, X4$ karakterisano je tranzicionom distribucijom predstavljenom lukovima iscrtanim punom linijom na slici 1. Tranzicione distribucije za svako stanje grade $N \times N$ tranzicionu matricu a_{ij} koja ostaje nepoznata sve dok se model ne „razradi". Ovde a_{ij} predstavlja verovatnoću prelaska od stanja i do stanja j u narednom koraku. Šta više, slučajna varijabla $y(t)$ predstavlja emisiju u vremenu $t(y(t) \in \{y1, y2, y3, y4, y5, y6\})$. Za svako stanje karakteristična je verovatnoća distribucije na mogućim emisijama $y1$ do $y6$. Verovatnoće rezultante – predstavljene tačkastim linijama na slici 1 – grade $N \times M$ emisionu matricu, kojom se definiše verovatnoća svakog rezultata u skladu sa skrivenim stanjem u modelu. Tako $b_i(k)$ predstavlja verovatnoću opažanja znaka yk kada se proces nalazi u stanju i. Šta više, dat je N-dimenzionalni vektor $\pi \in \{\pi1, .., \pi N\}$ s početnim vrednostima verovatnoće za svako stanje. Stoga se HMM može prikazati kao λ, gde $\lambda = (X, Y, a, b, \pi)$.



Slika 1: *Grafički prikaz strukture HMM*

Što se tiče naših pretpostavki u poglavlju 2, mi smo definisali CPIS of ≥ 6 kao jednak stanju „crveno". Pored toga, nismo dozvolili nikakav prelaz (tranziciju) iz stanja „zeleno" u stanje „crveno" i podesili smo sve vrednosti verovatnoće stanja „crveno" na 0, osim tranzicije ovog stanja u samo sebe. Početna verovatnoća za stanje „crveno" podešena je na 0. Nismo postavljali nikakve druge pretpostavke niti ograničenja. U skladu sa potrebama istraživanja, u vezi sa HMM javljaju se četiri problema/pitanja:

1. Kad se uzme u obzir skup posmatranih CPIS, O sa nizovima $o_1,.., o_L$ i HMM λ, kako ćemo prilagoditi parametre modela a, b i π da bismo maksimizovali $P(O|\lambda)$ (Od sada ćemo ovaj problem nazivati razrada). Da bismo rešili ovaj problem primenićemo Baum-Welch algoritam (Baum i dr., 1970).

2. Najverovatniji niz emisija CPIS koji počinje u bilo kom momentu u vremenu. Da bismo rešili ovaj problem, primenićemo terminski, povratni, algoritam (Rabiner, 1989) i iskoristićemo ovu informaciju za predviđanje toka bolesti i početka pneumonije.

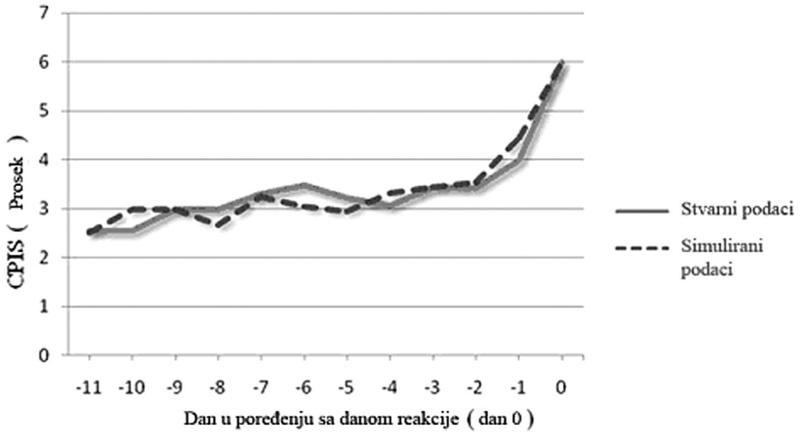
[1] **Ovo proizilazi it ograničenja temporalnih merenja nekih komponenata CPIS.**

3. Najverovatnije skriveno stanje u određenom momentu u vremenu, a time i buduća emisija. Ovaj problem ćemo rešiti primenom Viterbi algoritma (Forney, 1970) (Rabiner, 1989). Tako ćemo dobiti više informacija o toku same bolesti pneumonije.

4. Verovatnoća datog niza CPIS. Problem ćemo rešavati primenom terminskog algoritma pomoću koga ćemo sada klasifikovati nizove. Opis ovog skupa veoma poznatih algoritama dao je Rabiner, 1989.

4. Simulacija CPIS

Kad dokažemo da su stohastičke odlike CPIS vremenskih serija tačno predstavljene radnim HMM modelom, moći ćemo odmah da započnemo proces predviđanja i da brojimo tačna predviđanja. Da bismo shvatili koliko precizno HMM može da modeluje razvoj pneumonije, pretpostavićemo da će model moći dobro da predvidi tok samo ako može da simulira tok CPIS. Stoga smo razradili HMM onako kako je prikazano na slici 1, sa skupom slučajeva pneumonije. Na slici 2 prikazan je prosečni tok bolesti pacijenata obolelih od pneumonije i srednji tok višestrukih simulacija. Simulacija za jedan određeni tok počela je prvom slučajnom emisijom, a zaustavili bismo je ako bismo smo opazili CPIS ≥ 6. S obzirom na činjenicu da se dužine generisanih nizova razlikuju, svi nizovi su poređani oko dana reakcije. U poređenju sa serijama u realnom vremenu, simulirani nizovi imaju devijaciju u procentu od 6,95% prosečnog CPIS. Ovakav rezultat jasno pokazuje da je model prilagođen stohastičkim odlikama datih vremenskih serija.



Slika 2: *Simulirane i stvarne vremenske serije*

5. Postavka testa

Postavkom testa ćemo prilagoditi i simulirati tok bolesti pneumonije i odrediti verovatnoću oboljevanja od pneumonije u budućnosti. Ako se postigne ovakva funkcionalnost, sistem se može pretvoriti u svojevrsno „svetlo upozorenja". Prema tome, model može da posluži kao podrška za odlučivanje kad lekar treba da utvrdi dijagnozu. Ovakav test bi trebalo da predvidi manifestovanu pneumoniju – „crveno" stanje – upravo dan pre dana reakcije. Za sve ostale momente u vremenu predviđanje će biti „zeleno", „žuto" ili „narandžasto" stanje. Konstrukcija modela zavisi od medicinskih dokaza da će se pacijenti koji pokazuju visok nivo podložnosti pneumoniji zaraziti i oboleti od ove bolesti mnogo brže nego što je to uobičajeno (Oroszci, 2008). Model prikazuje ovaj koncept tako što primenjuje paradigmu slaganja kao što je prikazano na slici 3. Dva modula operišu u nizu: jedan razdvaja pacijente niskog i visokog rizika (klasifikacija), drugi pravi konkretnu prognozu i za jednu i za drugu grupu (predviđanje). Teorijski osnovi slaganja opisani su u odgovarajućoj literaturi (Wolpert, 1992). Slaganje (engl: *stacking*) predstavlja metod primene višestrukih nizova ili paralelnih modela da bi se dobila veća preciznost u predviđanju (Ting i Witten, 1997). Projektovanje modela predviđanja prijemčivosti vodi se sledećim zahtevima:

- Koristi se učenje pod nadzorom. Ovo zahteva da se koncept prijemčivosti razloži na veličine koje se mogu posmatrati. Kao prvu aproksimaciju koristili smo prvu pojavu pneumonije prema CRISP kao varijabli klase. Kao dodatak, može se primeniti skrivena varijabla koja proizilazi iz Modelovanja strukturne jednačine (SEM) (videti, na primer, Buncher i dr., 1991).

- Model treba da vrši predviđanja verovatnoće pomoću parametara koji omogućavaju da se njegove α i β greške lako prilagode. Učinak sistema će na kraju da odredi interakcija različitih delova modela.

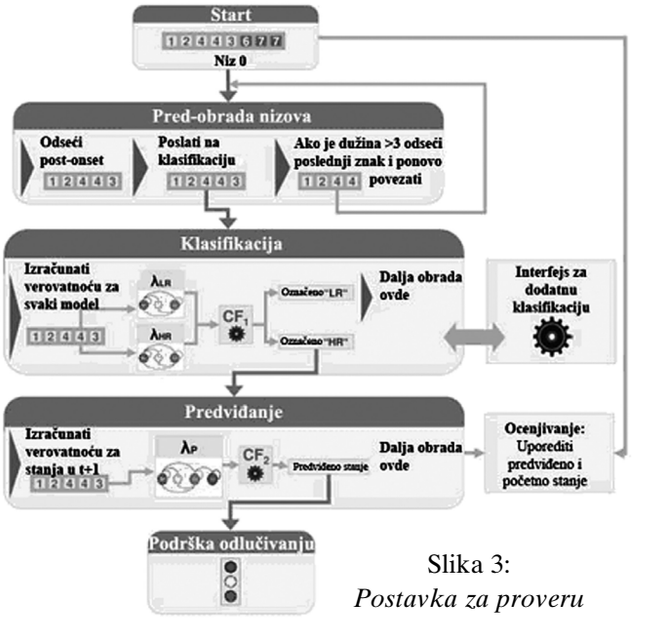
- Model mora da se bavi veoma neuravnoteženim uzorcima za učenje, pošto je grupa visoke prijemčivosti na bolest daleko manja od grupe pacijenata malog rizika.

Klasifikacija Razvijeni model predstavlja kombinaciju HMM i Bayesovog sistema. U delu klasifikacije koji je prikazan na slici 3, izvedena su dva HMM, jedan za pacijente visokog rizika (HR pacijenti - λ_{HR}) i jedan za ostale (LR pacijenti - λ_{LR}). Po Baum-Welch radnom modelu svaki model je prilagodio karakteristike slučajeva sa pneumonijom (λ_{HR}) i onih koji ovo oboljenje nemaju (λ_{LR}). U oba slučaja, vremenski nizovi koji se koriste za kalibraciju isključuju dan reakcije. Šta više, modeli HMM za modele klasifikacije i predviđanja se

neznatno razlikuju u broju stanjau skadu s činjenicom da stanje „crveno“ ne postoji u fazi pre početka u modelu klasifikacije. Da bi se klasifikovalo stanje pacijenta *o*, ono treba da se unese u oba kalibrirana HMM, a terminski algoritam će pokazati verovatnoću da se ovaj niz pojavi po Modelu λ_{HR} i λ_{LR} , odnosno ($P(o|\lambda_{LR})$ i $P(o|\lambda_{HR})$. Bayes-ova formula će pokazati *P(HRđo):*

$$P(\lambda_{HR} \mid o) = \frac{P(o \mid \lambda_{HR}) \cdot P_{HR}}{P(o \mid \lambda_{HR}) \cdot P_{HR} + P(o \mid \lambda_{LR}) \cdot P_{LR}} \qquad (1)$$

P_{LR} i P_{HR} pokazuju uobičajene a priori verovatnoće koje su uzete iz opšte statistike o zdravlju u ICU. Da bi rezultat bio validan, strukture (*X,Y,N,M*) i jednog i drugog HMM, i λ_{HR} i λ_{LR}, moraju da budu jednake. Pored toga, uvodi se prag ili faktor sigurnosti CF_1 (videti sliku 3). Time se prihvata klasifikacija *HR* ukoliko verovatnoća ($P(o|\lambda_{LR})$ dostigne određeni nivo. Ovim dobijamo mogućnost da odredimo minimalnu donju granicu za prihvatanje klasifikacije *HR*. Kada se fokusiramo na *HR* slučajeve u ovom radu, važi pravilo odlučivanja (koje će sa svoje strane uticati na α i β greške u ovoj fazi) da se slučaj označava kao *HR* ako njegova a posteriori verovatnoća $P(HR|o)$ pređe prag CF_1. Slučajevi klasifikovani kao *HR* proći će kroz klasifikator. Precizno govoreći, potreban je jedan analogni prediktor/procesor i za slučajeve *LR*. Pošto je rizik oboljenja od pneumonije u ovoj grani daleko manji, još nismo razradili model. U ovom slučaju korist od paradigme slaganja još jednom se potvrđuje, pošto će svaki korektno klasifikovan niz smanjiti broj β grešaka u narednim koracima. Vratićemo se problemu neuravnoteženih podataka u nizu. Pošto se svaki HMM posebno prati, ne postoje ograničenja za podjednako uravnotežene grupe sve dok postoji dovoljan broj podataka za praćenje. Ovo još jednom pokazuje koliko je HMM koristan u klasifikaciji.



Slika 3:
Postavka za proveru

Predviđanje Pošto niz o prođe kroz klasifikator i bude obeležen kao „HR“, model za predviđanje koji se sastoji od jednog HMM λ_P pravi prognozu na osnovu karakteristika o. Pošto smo rešili četiri problema iz poglavlja 3, lako ćemo dati prvu jednodnevnu prognozu od dana t do dana t + 1. Na osnovu vremenske serije signala koje smo pratili kod svakog pacijenta može se izračunati verovatnoća da skriveno stanje u t bude i. Ako poznajemo (skrivene) verovatnoće prenosa a_{ij}, mogu se izračunati mogućnosti za svako skriveno stanje na t + 1, što se onda može pretvoriti u verovatnoće za posmatranu emisiju na t + 1. Početak pneumonije je predviđen ako verovatnoća da će se dostići skriveno stanje „crveno“ ili posmatrana emisija ”CPIS > 6” pređe prag, t.j.,

- *P(”red”, t + 1) > CF_2* or
- *P(”CP IS > 6”, t + 1) > CF_2’*

Mogu se izvesti i sofisticiranija pravila, na primer: *P(”red”, t + 1) > CF_2”P (x, t + 1)* or for all other hidden states x and a relative threshold of CF_2

Našu diskusiju u ovom radu ograničićemo na prvi slučaj. Sistem se u principu može proširiti tako da prognoze pokrivaju i period duži od *t + 1*. Ako se uzme u obzir period inkubacije pneumonije koji traje dva ili tri dana, ovo nije od koristi za slučaj ove bolesti , ali jeste interesantno iz opšte perspektive. Da zaključimo, naš kompleksni model $\kappa = \kappa (\lambda_C , \lambda_N , \lambda_P , CF_1, CF_2)$ po definiciji sadrži sve pod-modele i parametre. Da bismo konačno izmerili kvalitet projekcije, razlažemo ceo skup podataka u skup za probu i skup za proveru. Skup za probu koristili smo da isprobamo modele klasifikacije i model za predviđanja na osnovu Baum-Welch algoritma. Sistem je obradio skup za proveru i rezultati predviđanja mogli su da se uporede sa stvarnim podacima[2] koje sistem ne poznaje. Prema slici 3, čitav proces funkcioniše na sledeći način:

1. Prvo, sistem određuje niz za testiranje o dužinom *T*, i ostavlja pred-fazu (o^* sa dužinom T^*). Vremenske serije pretpočetne dužine ≤ 3 ne uzimamo u obzir zato što kratke vremenske serije nisu od značaja.

2. Naravno, sistem za podršku u odlučivanju neće samo predvideti dan reakcije u nastanku pneumonije, već će i izbeći lažne pretpostavke koje prethode tom danu i uopšte u slučajevima gde pneumonije nema. Jedinstvena vremenska seri-

[2] **Što znači sa stvarnim stanjem u kome se pacijent nalazi**

ja pneumonije stoga takođe daje odsečke *(o*1, .., o*T* – 2*) izdvojene iz momenata (vremena) pre početka bolesti koji se onda korektno identifikuju kao serija „nema dana reakcije“ (povezati o^* sa korakom 1).

3. Klasifikacija „HR“ i „LR“. Ako se niz može označiti sa „HR“, predviđanje se nastavlja. U ovom trenutku mogu se uključiti i dodatni metodi i modeli, kako smo već naveli.

4. Predvideti narednu fazu prema λ_P i CF_2.

5. Uporediti predviđeno stanje sa stvarnim stanjem.

Fino podešavanje sistema Da bismo konačno započeli prognoziranje, model mora da reši još jedno suštinsko pitanje:

Koji je pravi način za tretiranje grešaka da bi se dobio optimalan rezultat? Model mora pravovremeno da predvidi pneumoniju i da istovremeno izbegne lažnu uzbunu pre tog trenutka, pokazujući greške α i β. Parametri CF_1 i CF_2 su osnovni parametri projekcije koji se mogu podešavati da bi se dobio „optimalni“ rezultat. Između tri tipa grešaka zapaža se jedan osnovni obrazac zamene:

- Greška kategorije 1 – lažni negativni: sistem nije uspeo da identifikuje dan reakcije

- Greška kategorije 2 – lažni pozitivni: sistem je identifikovao dan reakcije kod pacijenta obolelog od pneumonije u pogrešnom trenutku

- Greška kategorije 3 – lažni pozitivni: sistem je identifikovao dan reakcije kod pacijenta koji nije oboleo od pneumonije.

U medicinskoj praksi greške kategorije 1 smatraju se ozbiljnijim od grešaka kategorije 3, a obe ove vrste su daleko ozbiljnije od grešaka kategorije 2. Kliničkim jezikom rečeno, neke greške kategorije 3 mogu se smatrati lošom klasifikacijom, pošto se zamagljuje razlika između pneumonije i drugih vrsta plućnih bolesti kao što je bronhitis, a granica definisana kao CPIS ? 6 u stvarnosti postaje fazi granica. Da bismo rešili ovaj problem već smo uveli naša dva parametra (CF_1) i (CF_2). CF_1 i CF_2 sada se mogu primeniti tako da se postavka pomeri prema funkcionisanju koje „uspešnije izbegava lažne negativne“ prognoze, mada to obuhvata i veći broj nepotrebnih tretmana što se ovde ne smatra velikim problemom. Sada postoje dva moguća načina da se definišu CF_1 i CF_2. Prvo rešenje jeste da ordinirajući lekar definiše ova dva parametra kao fiksne konstante. Nedostatak ovog metoda jeste u tome da je definisanje prametara na način „crne kutije“ vrlo apstraktno i nije intuitivno pošto posledice nisu odmah vidljive. Druga mogućnost je da se parametri optimizuju u skladu sa datim radnim podacima i da se najbolji odnos u odnosu uspešno-greška. Stoga prvo treba operacionalizovati jedan optimalni odnos između uspešnog i greške. Da bismo rešili ovaj problem uvodimo ciljnu funkciju $F = F (CF_1, CF_2$ u kojoj, da bi se zadovoljile pretenzije korisnika, treba da se vrednuju 3 različite kvalitativne funkcije:

- Kvalitativna funkcija QF_1, koja predstavlja procenat tačno predviđenih dana reakcije u uslovima: model k i vrednosti CF_1 i CF_2.

- Kvalitativna funkcija QF_2, koja predstavlja procenat tačno predviđenih dana reakcije za slučajeve oboljenja od pneumonije u uslovima: κ, CF_1, CF_2.

- Kvalitativna funkcija QF_3, koja predstavlja procenat tačno predviđenih dana reakcije za slučajeve gde ne postoji oboljenje od pneumonije u uslovima: κ, CF_1, CF_2.

Jasno je da sve kvalitativne funkcije neposredno zavise od izbora CF_1 i CF_2. Na primer, ako se CF_1 podesi na 1, praktično nikada nećemo dobiti klasifikaciju „HR“ i stoga je dan reakcije skoro nemoguće predvideti. S druge strane, neće se pojaviti ni lažna prognoza. Pored toga, ciljnoj funkciji dodata su 3 parametra, P_1,P_2 i P_3, kao popravna mera za 3 kvalitativne funkcije ukoliko njihove vrednosti ne dostignu minimalni nivo. Ovo se može iskoristiti da se postavi niži nivo ograničenja za preciznost postavke.[3]

$$\begin{aligned} F(CF_1, CF_2) &= w_1 \cdot QF_1(k \mid CF_1, CF_2) - \\ & w_2 \cdot QF_2(k \mid CF_1, CF_2) - \\ & w_3 \cdot QF_3(k \mid CF_1, CF_2) - P_1 - P_2 - P_3 \end{aligned} \quad (2)$$

Ciljna funkcija F odražava razmene kad imamo veliki broj predviđenih dana reakcije i s druge strane veliki broj pogrešnih predviđanja.

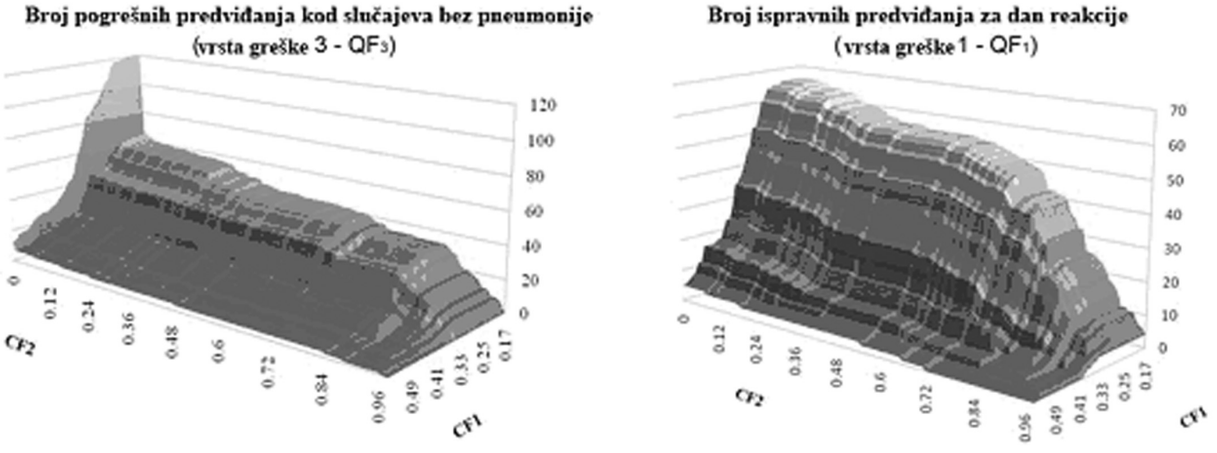
Sada nam je uz pomoć F lakše da operacionalizujemo potrebe korisnika kad definišemo zahteve. Rečima, korisnik može da definiše sledeće: „Sistem mora da identifikuje najmanje 40% svih dana reakcije (P_1), korektno predviđanje dana reakcije (QF_1 and w_1) je dvostruko važnije nego izbegavanje pogrešnih prog-

[3] **Na primer, kao popravna mera kada se u rezultatu pojavi preveliki broj pogrešnih predviđanja.**

noza kod pacijenata koji nisu oboleli od pneumonije (QF_3 and w_3). U ovom slučaju, na primer, ima smisla vrednovati izbegavanje greške kategorije 1 (w_1 za QF_1) koja je višeg reda, da bi se sprečio izostanak lečenja. Definisanjem vrednosti w_1,.., w_3, primenom ciljne funkcije F dobijamo određenu ciljnu vredost za svaku kombinaciju CF_1 i CF_2. Stoga sistem može da proizvede skalarne optimalne kombinacije CF_1 i CF_2 u skladu sa definicijama ovih vrednosti. U ovom radu ne možemo da ispitujemo definicije vrednosti za svaki parametar pošto je to složen proces utvrđivanja u medicini. Umesto da predstavimo jedno optimalno rešenje, predstavićemo odnos greška-uspeh za svaku postavku parametara i neke rezultate kao primer.

6. Eksperimentalni rezultati

Na slici 4 predstavljamo efekte α i β greške u odnosu na broj tačno predviđenih dana reakcije i broj pogrešno predviđenih dana reakcije (u grupi slučajeva gde nema pneumonije) u zavisnosti od kombinacije vrednosti CF_1 i CF_2. Očigledno, oblik pokazuje izvesnu korelaciju koja odslikava razmene α – β grešaka. Kako se broj tačno predviđenih dana reakcije smanjuje, smanjuje se i broj pogrešnih predviđanja. Bez obzira na to, oblici ukazuju na razlike koje omogućavaju da se izvrši prilagođavanje ova dva parametra. Opseg niskih vrednosti CF_2 posebno pokazuje neuobičajeno veliki rast broja grešaka koje se javljaju uporedo sa manje ili više stabilnim brojem tačnih predviđanja.



Slika 4: *Površi optimizacije QF_3 i QF_1*

Na tabeli 2 prikazani su rezultati u svakoj kategoriji greške. U prvom delu pokazujemo do koje mere sistem može da odredi da li vremenska serija pripada slučaju sa ili bez pneumonije, a na osnovu različitih radnih metoda. Sa stvarnom pozitivnom stopom od oko 82-83% sistem jeste u stanju da odredi vremenske serije kao stanje pneumonije ili stanje bez pneumonije. U drugom delu koristili smo skup od 3 različita rasporeda vrednosti da pokažemo način na koji naš model radi i razmeni α – β grešaka. Kako raspored vrednosti 3 (dobijen od ordinirajućeg lekara) pokazuje porast broja prognoza dana reakcije, preciznost se može žrtvovati u korist smanjivanja broja grešaka tipa 3. Kao referentnu vrednost koristili smo dve laičke strategije predviđanja i poredili rezultate. Prvi laički metod predviđa dan reakcije u t + 1 ukoliko se vrednost CPIS „5“ postigne u t. Drugi izračunava krivu vremenskih serija na osnovu vrednosti t i t – 1. Kao što se vidi na tabeli 2, laički metod 1 predstavlja u stvari veoma puozdanu prognozu koja u velikoj meri utiče na način na koji se stvara zbir CPIS. Prednost našeg metoda jeste u tome da se (putem izbora praga CF_1 i CF_2) može podesiti vrednost α, a ona će takođe odrediti vrednost β (i obrnuto). Šta više, ovaj metod pokazuje da je prilagođavanje α grešaka moguće a da istovremeno imamo značajnu stabilnu β grešku. Stoga se može istraživati konkretna razmena između ova dva tipa greške.

Tabela 2: *Eksperimentalni rezultati*

Model klasifikacije			
Radni model		Slučajevi pneumonije (N=79)	Slučajevi bez pneumonije (N=285)
Standardni Baum-Welch radni tip		82%	69,8%
Poboljšani radni tip (Genetski algoritam)		83%	73,5%
Ukupni rezultati			
Metod	Tačno predviđen dan reakcije	Kategorija greške 2: prerano predviđen dan reakcije	Kategorija greške 3: pogrešno predviđen dan reakcije
κ postavka vrednosti 1	41,6%	6,9%	6,5%
κ postavka vrednosti 2	63,8%	11,6%	9,3%
κ postavka vrednosti 3	43,2%	20,1%	6,5%
Trivijalni metod 1 ($y5\ t \rightarrow y6\ t+1$)	44,3%	21.7%	13,8%
Trivijalni metod 2 (kriva)	31,6%	15.2%	10,4%

7. Zaključak

Test koji smo primenili u ovom radu može se smatrati novim pristupom u obradi i predviđanju podataka u medicini. Ipak, ovo istraživanje predstavlja samo prvi pokušaj da se pneumonija analizira primenom HMM i on ima neka ograničenja. Izračunavanja smo vršili na skupu podataka prikupljenih u periodu od dve godine, na samo 79 (slučajevi pneumonije) i 285 (slučajevi bez pneumonije) vremenskih serija tako da bi ih trebalo uporediti i sa podacima za druge godine da bi se procenio kvalitet. Pored toga, podaci su uzeti iz jednog ICU tako da nije jasno da li su rezultati posledica i nekog lokalnog uticaja i da li neka druga ICU možda imaju drugačije rezultate. Da bismo predvideli VAP koncentrisali smo se na obradu vremenskih serija slučajeva pneumonije. Prema modelu klasifikacije, u ovom trenutku nismo išli dalje i obrađivali nizove označene kao „LR“. Da bi se uspostavio holistički sistem potrebno je dalje raditi u ovoj oblasti. Pored toga, možda će biti potrebno i da proverimo i istražimo implikacije koje su dovele do strukture HMM (broj stanja itd.). Ciljna funkcija koju smo primenili da bismo dobili optimalni rezultat može se proširiti kako bi obuhvatila i ekonomska pitanja kao što su konkretni troškovi lečenja. Druga korist od ovog istraživanja jeste što je potvrdilo znanja koja smo imali o pneumoniji. Prema našim rezultatima, potvrđena je pretpostavka da je za pneumoniju karakterističan kratak period inkubacije. S druge strane, ovi rezultati imaju neke slabe strane. Ako posmatramo okvir CPIS od jednog dana, naša predviđanja će biti usko ograničena na ovaj kontekst. Dobro bi došlo da se posmatra širi vremenski okvir. Isto tako, sistem se zasniva na veoma razrađenoj pred-obradi podataka iz dvogodišnjeg perioda. Nažalost, ovi podaci su još uvek u velikoj meri nedovoljni zbog problema koje smo naveli u poglavlju 2. Stoga jedan integrisani sistem za rano otkrivanje mora da se zasniva na holističkom, a priori uklapanju u infrastrukturu podataka kojima bolnica raspolaže u realnom vremenu. Ako već nije realizovana, predupotrebna faza će zahtevati mnogo resursa. Instaliranje ovakvog sistema i dugo traje i skupo je, ali se zato može koristiti na više načina i u različitim situacijama. Ukoliko takvi sistemi postoje, sama prognoza pneumonije zahteva veoma malo resursa u svakodnevnom funkcionisanju. Ona se može i treba da se uključi u pacijentov „karton“ u kome se nalazi pregled istorije bolesti pacijenta i koji predstavlja koristan instrument za ordinirajuće lekare. Očigledno je da se sitem proširio na nekoliko mesta. Sistem se može proceniti i na primeru drugih oboljenja koja imaju duži period inkubacije. Na nivou klasifikacije mogu se primeniti i drugi metodi, na primer, Bayesove mreže. Pošto su se drugi koncepti predispozicije pokazali kao veoma moćni (Oroszci, 2008), trebalo bi ispitati i druge a priori metode, na primer, SEM. U okviru strukture slaganja u našem sistemu ovo se može postići bez problema.

Da zaključimo, pokazali smo da su metodi za izdvajanje podataka veoma uspešni i omogućavaju da se mnogo sazna iz uskladištenih medicinskih podataka. Svakako da se potpuno automatizovano rešenje „kao iz kutije“ ne može dobiti. Ipak, sistem pokazuje kako primena slaganjem u različitim metodima može da pospeši otkrivanje potencijalnih oboljenja skrivenih u podacima u datotekama.

LITERATURA

[1] Baum LE., Petrie T., Soules G., Weiss N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The Annals of Mathematical Statistics 41(1):164–171, URL http://www.jstor.org/stable/2239727

[2] Buncher C.R., Succop P.A., Dietrich K.N. (1991) Structural equation modeling in environmental risk assessment. Environ Health Perspect 90:209–213

[3] Bystroff C., Krogh A. (2008) Hidden markov models for prediction of protein features. In: Protein Structure Prediction, Humana Press, Methods in Molecular Biology, vol 413, pp 173–198, DOI 10.1007/978-1-59745-574-9 7, URL http://www.springerlink.com/content/g4111p42750174r2/

[4] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R. (2000) Crisp-dm 1.0 step-by-step data mining guide. Tech. rep., The CRISP-DM consortium, URL http://www.crisp-dm.org/CRISPWP-0800.pdf

[5] Forney G.D. (1973) The viterbi algorithm. Proceedings of the IEEE 61(3):268–278, URL http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber=1450960

[6] Gascuel O., Moret B.M.E. (eds) (2001) Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 28-31, 2001, Proceedings, Lecture Notes in Computer Science, vol 2149, Springer

[7] Heyland D.K., Cook D.J., Griffith L., Keenan S.P., Brun-Buisson C. (1999) The attributable morbidity and mortality of ventilator-associated pneumonia in the critically ill patient. The Canadian Critical Trials Group. Am J Respir Crit Care Med 159:1249–1256.

[8] Manning C.D., Schütze H. (2005) Foundations of statistical natural language processing, 8th edn. MIT Press, Cambridge, Mass.

[9] NN (2005) Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med 171:388–416

[10] Oroszi F. (2008) Einsatz von Data Mining Verfahren auf medizinischen Daten - Anwendungsfall Pneumonie-Früherkennung. PhD thesis, Friedrich-Schiller-Universität Jena

[11] Pugin J., Auckenthaler R., Mili N., Janssens J.P., Lew P.D., Suter P.M. (1991) Diagnosis of ventilator-associated pneumonia by bacteriologic analysis of bronchoscopic and nonbronchoscopic ”blind” bronchoalveolar lavage fluid. Am Rev Respir Dis 143:1121– 1129

[12] Rabiner L.R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE, pp 257–286

[13] Rabiner, L. R., and Juang, B. H. (1986), An Introduction to Hidden Markov Models, IEEE Acousfics, Speech & Signal Processing ,Magazine, 3, 1-16.

[14] Rea-Neto A., Youssef N.C., Tuche F., Brunkhorst F., Ranieri V.M., Reinhart K., Sakr Y. (2008) Diagnosis of ventilator-associated pneumonia: a systematic review of the literature. Crit Care 12:R56

[15] Tejerina E., Frutos-Vivar F., Restrepo M.I., Anzueto A., Abroug F., Palizas F., Gonzalez M., DÉmpaire G., Apezteguia C., Esteban A. (2006) Incidence, risk factors, and outcome of ventilator-associated pneumonia. J Crit Care 21:56–65

[16] Ting K.M., Witten I.H. (1997) Stacked generalization: when does it work. In: in Procs. International Joint Conference on Artificial Intelligence, Morgan Kaufmann, pp 866– 871

[17] Wolpert D. (1992) Stacked generalization. Neural Networks 5:241–259