Diana Sekulić i Tijana Aleksić

Klasifikacija raka analizom mikronizova

Cilj ovog rada je upoređivanje tačnosti tri klasifikatora pri klasifikaciji pet različitih tipova raka. Tipovi raka su klasifikovani na osnovu potpisa koje formiraju pojedinačne supstitucije unutar čovekovog genoma. Na osnovu skupa podataka jednog pacijenta, koji sadrži informacije o mutacijama i njihovim pozicijama u genomu, grupišu se karakteristični potpisi za svaki tip raka, i potom se klasifikuju. Poređeni su klasifikator maksimalne margine, logistička regresija i random forest algoritam za klasifikaciju. Klasifikacija je vršenja na bazi podataka koja sadži 1660 uzoraka 5 različitih tipova raka (rak dojke, rak debelog creva, rak jetre, rak pluća i rak grlića materice). Tačnost klasifikatora maksimalne margine je 76.8% sa preciznošću od 79.0%, logistička regresija je davala tačnost od 88.8% sa preciznošću od 90.7%, dok je random forest algoritam davao tačnost od 84.8% sa preciznošću od 84.7%. Rezultati su uporedivi sa rezultatima referentnih radova i istog su ranga tačnosti za dati broj mutacija (50).

Diana Sekulić (1999), Novi Sad, Ćirila i Metodija 122, učenica 4. razreda Gimnazije „J. Jovanović Zmaj" u Novom Sadu

Tijana Aleksić (2000), Kragujevac, Dragoljuba Božovića Žuće 7, učenica 4. razreda Prve kragujevačke gimnazije

MENTORI:

Gavrilo Andrić, Seven Bridges Genomics, Beograd

Natalija Krsmanović, Majkrosoft razvojni centar Srbije

Pavle Šoškić, student Elektrotehničkog fakulteta Univerziteta u Beogradu

Uvod

Biološka osnova

Rak predstavlja grupu bolesti koja uključuje abnormalni rast ćelija sa potencijalnom mogućnošću da se poremeáj proširi na ostale delove tela. Karakterišu ga nesvrhovitost, autonomnost, kao i činjenica da nastavlja rast i nakon prestanka funkcionisanja samog uzroka tumora. Tumorske bolesti često mogu dovesti do smrti. One se manifestuju putem tumorskih izraslina koje nastaju prilikom gubljenja normalne regulacije kontrolnih mehanizama rasta ćelije i od nezaraženih ćelija razlikuju se strukturno i funkcionalno. Postoje dve vrste tumora: maligni i benigni, a razlika između njih se ogleda u tome što benigni tumori ne interaguju sa okolnim organima, dok se maligni šire po čitavom telu domaćina (Hanahan i Weinberg 2000). Uzročnici raka mogu biti kako unutrašnji tako i spoljašni. Spoljašnji faktori su faktori poput UV zračenja koje u najvećoj meri izaziva rak kože,

konzumiranja duvana koje izaziva rak pluća, ali ima udeo i u stvaranju mnogih drugih vrsta raka, i mnogi drugi uzročnici (Ferlay *et al*. 2013). Pored prethodno pomenutih uzročnika, rak se može naslediti od roditelja (unutrašnji uzročnici raka), što se dešava u 5-10% slučajeva (ACS 2018). Svi ovi uzročnici dovode do promene ćelijske DNK usled mutacija.

DNK (dezoksiribonukleinska kiselina) jeste nukleinska kiselina koja ima razne uloge u organizmima, od kojih je uloga za prenos informacija za sintezu proteina najvažnija. Sastoji se od nukleotida kojeg čine pentozni šećer dezoksiriboza, fosfatna grupa i azotne baze. Azotne baze se na osnovu hemijskog sastava dele na dve grupe: purinske i pirimidinske baze. Purinske baze su adenin (A) i guanin (G). Pirimidinske baze su citozin (C) i timin (T). Ove baze se u molekulu DNK vezuju po principu komplementarnosti, adenin i citozin, guanin i timin.

Mutacije do kojih dolazi u organizmu mogu se podeliti na tri karakteristične promene:: supstitucije, delecije i insercije. Supstitucija je zamena jednog nukleotida i njegovog para u komplementarnom lancu. Supstitucije se dele na:

1) Tranzicije – pri kojima se jedan pirimidinski nukleotid zamenjuje drugim pirimidinskim nukleotidom (C u T ili T u C) ili jedan purinski drugim purinskim nukleotidom (A u G ili G u A); dakle, postoji 4 tipa tranzicija;

2) Transverzije – pri kojima se pirimidinski nukleotid zamenjuje purinskim (A > C, G > C, A > T ili G > T) ili obrnuto iz čega proizlazi da postoji 8 tipova transverzija.

Delecije su promene kod kojih dolazi do izostavljanja jednog ili više nukleotida. Insercije su promene kod kojih dolazi do umetanja jednog ili više nukleotida na mestima gde to nije predviđeno.

Klasifikacija

Na osnovu mutacija, koje karakterišu sam tip promene, kao i mesta u genomu čoveka na kojima su se javile, moguće je odrediti skupove mutacija koji najbliže određuju i karakterišu specifični tip raka, te ih je moguće i klasifikovati (Alexandrov *et al*. 2013).

Postoji veliki broj istraživanja na temu klasifikacije raka. Ranija istraživanja zasnivala su se na analizi ekspresije gena, dok se trenutno radi na tome da se klasifikacija vrši generisanjem potpisa na osnovu mutacija koje se javljaju u genomu čoveka, i koji karakterišu određenu vrstu raka. U novije vreme pojavila su se istraživanja koja ukazuju na to da mutacija određenih gena u čovekovom genomu generišu određene potpise, i time ukazuju na određeni tip raka (Alexandrov *et al*. 2013).

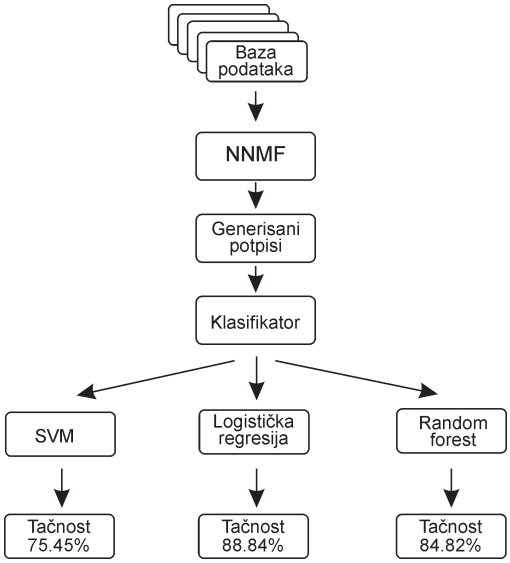
Na osnovu ranijih istraživanja u polju bioinformatike, izabrani su algoritmi koji će se koristiti za klasifikaciju. Najzastupljeniji algoritam koji se koristi jeste SVM (klasifikator maksimalne margine, eng. support vector machine) koji daje visok procenat tačnosti klasifikacije. Pored SVM klasifikatora koriste se i drugi linearni klasifikatori poput logističke regresije. Tokom poslednjih godina u sferi bioinformatike postaju popularni i

nelinearni klasifikatori, a posebno se iz njih izdvaja Random forest (Soh *et al*. 2017). Pored nelinearnih klasifikatora takođe se razvio i način klasifikovanja pomoću deep learning algoritama koji trenutno imaju najveću uspešnost u oblasti bioinformatike. Problem kod korišćenja deep learning algoritama jeste to što je potrebna značajno veća količina podataka kako bi se postigla željena tačnost. S obzirom da se na osnovu DNK lanca čoveka može dobiti veliki broj informacija o njemu, zbog privatnosti pacijenata, kao i zbog visoke cene sekvenciranje DNK jako mali broj podataka je dostupan za korišćenje i rad. Iz navedenih razloga deep learning algoritmi nisu često korišćeni. U istraživanjima koja imaju istu tematiku kao i sprovedeno istraživanje, najveću uspešnost imao je SVM algoritam za klasifikaciju.

Na osnovu referentnih radova (Soh *et al*. 2017) i karakteristika odabranih klasifikatora, očekivano je da SVM algoritam za klasifikaciju ima najveći tačnost pri klasifikaciji, zatim logistička regresija, dok se za random forest očekuje da ima najmanju tačnost pri klasifikaciji. Takođe, pretpostavlja se da će se sa povećanjem minimalnog broja mutacija po pacijentu povećavati i tačnost svakog algoritma posebno. Ako je minimalan broj mutacija koji ulazi u algoritam za generisanje potpisa mali, potpisi će sadržati manje informacija i neće se razlikovati dovoljno, tj. postojaće veće sličnosti između potpisa različitih tipova kancera. S druge strane, ako povećamo broj minimalnih mutacija, potpisi će se generisati jasnije.

Metod

Radi klasifikacije različitih tipova raka, najpre treba pripremiti bazu i iz nje izdvojiti sve potrebne podatke. Nakon toga, primenom algoritma NNMF (nenegativna faktorizacija matrica, eng. non-negative matrix factorization), generišu se potpisi na osnovu kojih se kasnije raspoznaju



Slika 1. Blok šema klasifikacije

Figure 1. Block sheme of classification

tipovi kancera. Poslednji korak jeste implementacija samih klasifikatora i evaluacija njihovih performansi (slika 1).

Priprema podatka koje klasifikator koristi podrazumeva prikupljanje svih podataka o pacijentu potrebnih za klasifikaciju, kao što su informacija o tipu mutacije i informacija o poziciji na kojoj je došlo do mutacije. Za istraživanje su korišćeni podaci pacijenata kod kojih su se pojavile karakteristične supstipstitucije pojedinačnih baza u genomu. Uticaj na generisanje potpisa ima i najbliža okolina same mutacije koja se može dobiti pronalaženjem mesta mutacije u referentnom genomu. Tako pripremljena baza se može koristiti za dalja istraživanja.

Baza. Podaci koji su korišćeni pri analizi i klasifikaciji tipova raka sastojali su se od 1660 uzoraka kancera 5 različitih tipova. Posmatrani tipovi raka su: rak dojke (BRCA) sa 764 različitih uzoraka, rak debelog creva (COAD) sa 260 različitih uzoraka, rak jetre (LINC) sa 213 različitih uzoraka, rak pluća (LUSC) sa 177 različitih uzoraka i rak grlića materice (UCEC) sa 246 različitih uzoraka.

Za svakog pacijenta u bazi postoje podaci o tipu mutacije i podatak o tačnoj poziciji na kojoj se mutacija pojavila u hromozomu. Mogu se javiti sva tri tipa mutacija. Za potrebe ovog istraživanja posmatrane su samo supstitucije radi pojednostavljivanja modela potpisa, a kako su supstitucije najzastupljeniji vid mutacija i na osnovu istraživanja u radu Ludmila Aleksandrova i saradnika (Alexandrov *et al*. 2013), one predstavljaju i dovoljan uslov za klasifikaciju.

Generisanje potpisa. Različiti procesi mutacija u ljudskom organizmu često dovode do različitih kombinacija mutacionih tipova koji se nazivaju potpisi. Potpise određuju skupovi različitih supstitucija, delecija i insercija koji se javljaju u organizmu čoveka.

U radu su posmatrane isključivo supstitucije, i na osnovu njih su generisani svi potpisi. Analizom baze pacijenata uočeno je da se delecije i insercije ne javljaju kod svih pacijenata, a i kada se pojave, pojave se u razmeri 1:50 naspram supstitucija, tako da se mogu zanemariti. Potpisi se generišu posebno za svaku bolest i ne preklapaju se međusobno. Da bi se generisao potpis potrebno je razvrstati sve supstitucije u 6 glavnih klasa na osnovu toga koja baza je zamenila koju. Postoji 6 glavnih klasa supstitucije (C > A, C > G, C > T, T > A, T > C, T > G) (Alexandrov *et al*. 2013). Klasa C > A na biološkom nivou označava to da se na mestu gde bi se na osnovu referentnog genoma nalazila baza C u mutiranom genomu pacijenta nalazi baza A. Analogno važi i za ostale klase. Nakon što se supstitucije razvstaju u 6 osnovnih klasa, vrši se dalje razvrstavanje po podklasama. Na osnovu podatka o poziciji unutar hromozoma na kojoj se desila supstitucija, u referentnom genomu se pronalaze podaci o tome koje baze se nalaze u okolini same mutacije. Za okolinu je uzeta po jedna baza sa obe strane mutacije. Prema tome, potpis se sastoji od podataka o supstitucijama kod pacijenta koje su razvrstane u 96 podklasa vezanih za svih 6 klasa. Za generisanje potpisa koristi se standardna metoda NNMF.

NNMF je grupa algoritama koja se koristi za aproksimaciju neke matrice. Koristi se za smanjenje matrica nad kojim se vrši ispitivanje, te

samim tim i lakšim rad nad njima, kao i u poboljšanju brzine izvršavanja. Faktorizacija se vrši na dve matrice manjih dimenzija. U istraživanju, ulazna matrica (V) je matrica koja sadrži podatke o broju supstitucija u svakoj klasi kod svakog pacijenta, dok matrice koje se dobijaju faktorizacijom sadrže podatke o tome koliko je koja signatura zastupljena kod svakog pacijenta (W) i druga matrica (H) koja predstavlja zastupljenost svake klase supstitucija u svakom potpisu:

$$[V]_{n\times 96} = [W]_{n\times 22} \times [H]_{22\times 96} \quad (1)$$

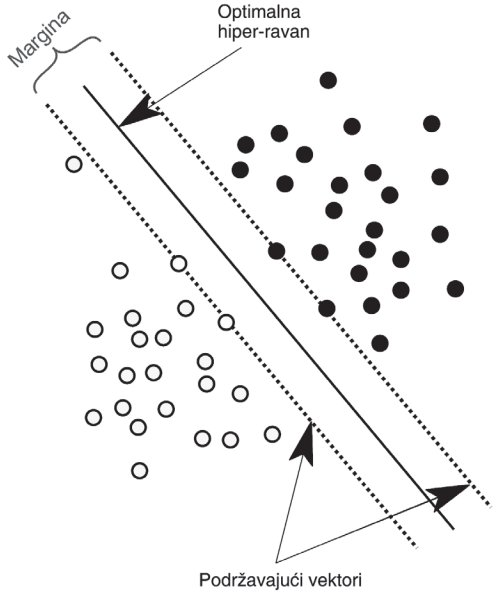
gde n predstavlja broj pacijenata.

Kada su potpisi za svaku bolest generisani, matrice potpisa se prosleđuju klasifikatoru. Uzorci koji se prosleđuju klasifikatoru jesu matrice W (formula 1), koje predstavljaju sve bolesti generisane faktorizacijom.

Klasifikacija. Implementirana su tri klasifikatora koji se koriste za klasifikaciju u više klasa, kako bi se uzorci tipova tumora klasifikovali u 5 različitih klasa kancera. Implementirani klasifikatori su: SVM, logistička regresija i šuma nasumične odluke. Iako deep learning algoritmi daju bolje rezultate pri klasifikaciji, za potrebe ovog istraživanja su odabrani jednostavniji klasifikatori. Baza sa kojom je rađeno imala je podatke od 1660 pacijenata, i upravo je ograničenost baze uslovila odabir klasifikatora.

SVM je zasnovan na principu vektorskih prostora. To je linearni klasifikator koji funkcioniše na principu pronalaženja margine koja ima najveće moguće ortogonalno rastojanje od podržavajućih vektora klasa koje treba da se razdvoje (slika 2), a pri tome zadovolji uslov da se što veći broj podataka iz iste klase nalazi sa iste strane margine. Model klasifikatora je formula, što znači da se klasa ulaznog podatka računa.

Baza je podeljenja na dva seta: za validaciju i za treniranje. Podaci za treniranje činili su 70% ukupne baze dok su podaci za validaciju činili 30%.

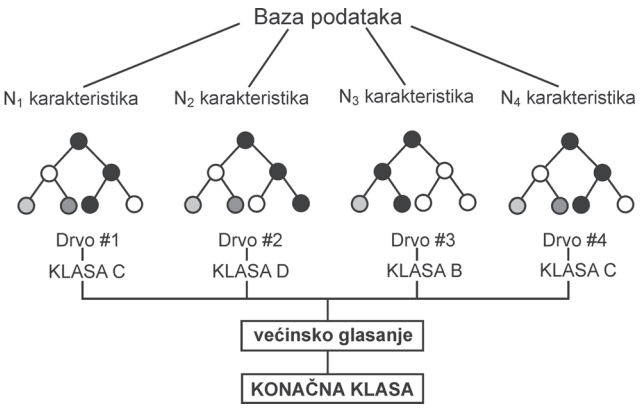


Slika 2. Klasifikator maksimalne margine

Figure 2. SVM classifier

Kao ulazni set se prosleđuju trening primeri kojima je pridružena klasa kojoj pripadaju, na osnovu kojih klasifikator fituje optimalnu marginu. Potom klasifikator za svaki od test-primera koji mu se prosleđuje predviđa klasu kojoj bi pripadao. Razlog zbog čega je pomenuti klasifikator primenljiv za ovo istraživanje jeste taj što je njegov rad predviđen za rad nad podacima čije dimenzije prevazilaze par stotina kB.

Logistička regresija je jedan od modela linearne regresije u kome zavisna promenljiva uzima samo dve vrednosti. To je statistička metoda koja se koristi pri binarnim klasifikacijama tako što svi podaci koji uzimaju vrednosti do određene granice pripadaju prvoj, a ostali drugoj klasi. Logistička regresija je specijalizovana forma regresije koja je formulisana da predvidi i objasni kategoričke varijable. Sistem funkcionisanja treniranja i validacije identičan je kao kod klasifikatora maksimalne margine, izuzev oblika podataka koji se dobijanju po završetku trening seta. Logistička regresija je primenljiva za ovo istraživanje jer su varijable koje se koriste kategoričke i linearno nezavisne.



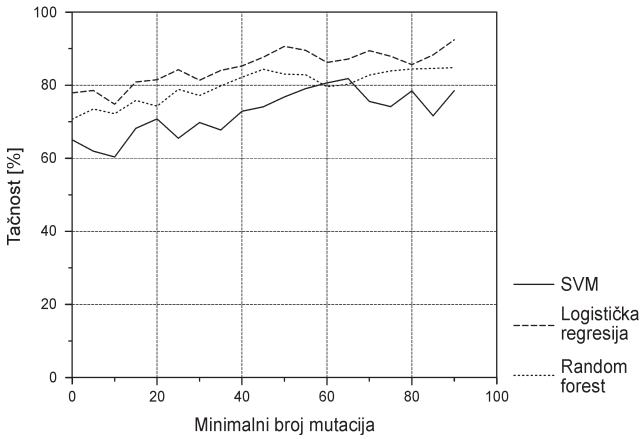
Slika 3. Klasifikacija klasifikatorom random forest

Figure 3. Random forest classifier

Random forest (slika 3) je specijalizovan način klasifikacije koji funkcioniše na principu stabla odluke. Stablo odluke (engl. decision tree) je stuktura nalik binarnim stablima, gde svaki čvor predstavlja test na određenu karakteristiku, dok listovi stabla predstavljaju konačnu odluku stabla. Random forest se sastoji od više različitih stabala odluka koji nezavisno jedni od drugih klasifikuju test primer. Konačna odluka random foresta se formira kada se usrednje odlučene vrednosti svih stabala, i uzorak se svrstava u onu klasu za koju je najviše stabala glasalo.

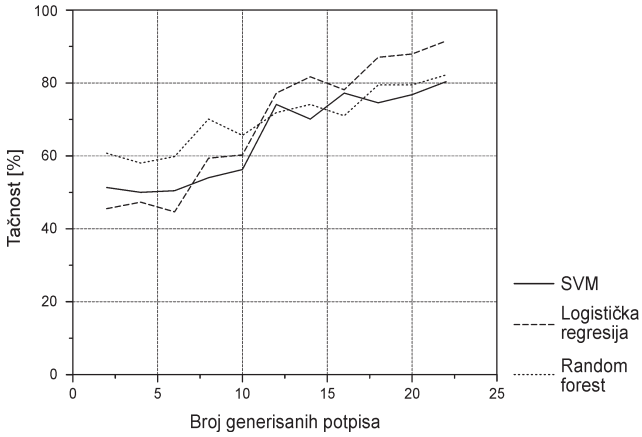
Rezultati

Tačnost svakog klasifikatora posmatrana je u zavisnosti od dva različita parametra. Prvi parametar koji je variran je minimalni broj supstitucija po svakom pacijentu za koga je vršena klasifikacija raka, dok je drugi parametar broj potpisa koji se generiše prilikom nenegativne faktorizacije.



Slika 4. Tačnost klasifikacije u zavisnosti od izabranog minimalnog broja mutacija (supstitucija)

Figure 4. Accuracy compared to the minimal number of mutations



Slika 5. Tačnost klasifikacije u zavisnosti od broja generisanih potpisa

Figure 5. Accuracy compared to the number of signatures that were generated

Na slici 4 prikazana je zavisnost tačnosti od minimalnog broja mutacija koje su uzimane u obzir. Promena minimalnog broja mutacija je uzimana od 0 do 90 sa korakom 5. Na slici 5 prikazana je promena tačnosti klasifikatora u odnosu na promenu generisanih potpisa. Broj potpisa je menjan od 2 do 22 sa korakom 2.

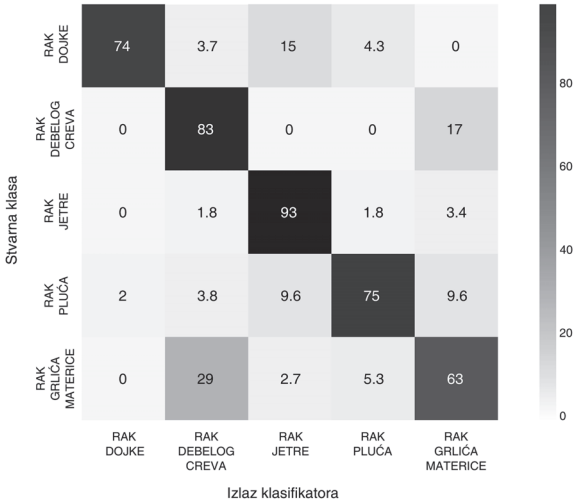
Baza podataka sadži ukupno 1660 pacijenata, što predstavlja znatno manju bazu od referentnih radova sa istom tematikom (primera radi jedan od radova ima 6 puta veću bazu). Kod ovako malih baza podataka potrebno je izabrati odgovarajuće parametre za evaluaciju klasifikatora da ne bi došlo do pojave preobučavanja (eng. overfiting) pri klasifikaciji. Kako bi se izbeglo da se radi sa količinom podataka koja će lako dovesti do overfitinga, ponovljenim merenjem se došlo do zaključka da 50 mutacija po pacijentu ostavlja zadovoljavajući broj pacijenata u bazi kako bi se klasifikatori trenirali, testirali i na kraju rezultati uporedili. Ukupna tačnost klasifikatora određena je pri sedećim parametrima: za minimalan broj mutacija je uzeto 50, a generisana su 22 potpisa. SVM klasifikator je imao tačnost od 76.8% sa preciznošću 79.0%, logistička regresija je imala tačnost od 88.8% sa preciznošću 90.7%, dok je random forest imao tačnost od

84.8% sa preciznošću 84.7%. Tačnost prikazuje udeo tačno klasifikovanih uzoraka od svih uzoraka koji se klasifikuju. Ona sama nije dovoljno informativna za evaluaciju klasifikatora, jer ne ukazuje na to koliko je svaki tip kancera tačno klasifikovan, pa se uz tačnost posmatra i preciznost. Preciznost prikazuje udeo tačno klasifikovanih pozitivnih odgovora u odnosu na sve odgovore koji su klasifikovani kao pozitivni. Iz razloga što tačnost, a isto tako i preciznost, ne uzimaju u obzir i tačne negativne odgovore, one nisu uvek dovoljne da bi se klasifikatori uporedili i da bi se odredio najbolji za dati problem. Iz tog razloga se uvodi treća ocena klasifikatora, F1-skor. F1-skor uzima u obzir i tačne pozitivne i tačne negativne odgovore, i na taj način predstavlja otežinjenu vrednost preciznosti i odziva (engl. recall – udeo tačnih pozitvnih u ukupnom broju tačnih dgvora) i omogućava pouzdanije poređenje. U tabeli 1 prikazane su sve tri evaluacije klasifikatora.

Tabela 1. Evaluacija klasifikatora

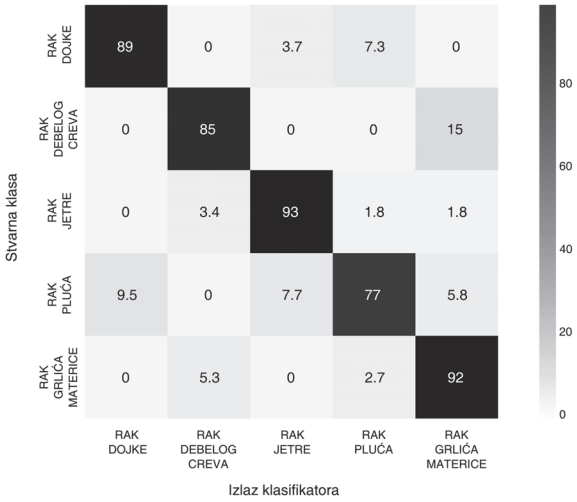
	SVM	Logistička regresija	Random forest
Tačnost	76.78%	88.84%	84.82%
Preciznost	79.05%	90.71%	84.66%
F1 skor	76.22%	90.59%	82.39%

Na slikama 6–8, prikazane su matrice konfuzije svakog klasifikatora (klasifikator maksimalne margine, logistička regresija i random forest). Matrice konfuzije su generisane za 22 generisana potpisa i pri minimalnom broju mutacija 50 za svakog pacijenta. Svaki red matrice predstavja stvarnu klasu podataka, dok kolone predstavljaju izlaz klasifikatora. Brojevi u poljima matrice predstavljaju procentualnu zastupljenost pripadnika neke klase kojima je dodeljena određena klasa (tj. procenat true positive i false positive).



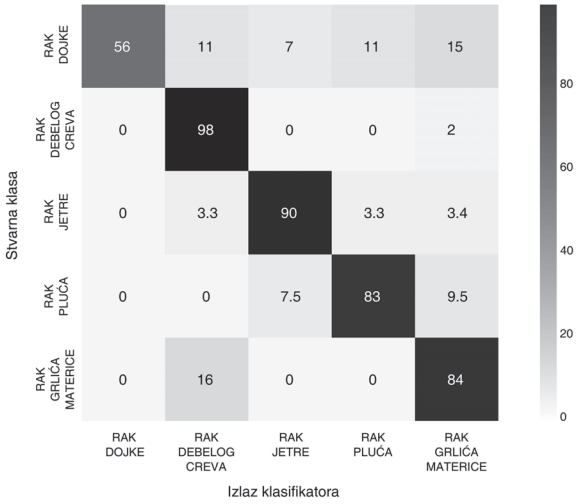
Slika 6. Matrica konfuzije klasifikatora maksimalne margine

Figure 6. Confusion matrix of SVM
From lefto to right/top to bottom:
breast cancer, colon cancer, liver cancer, lung cancer, cervical cancer



Slika 7. Matrica konfuzije logističke regresije

Figure 7. Confusion matrix of logistic regression
From lefto to right/top to bottom:
breast cancer, colon cancer, liver cancer, lung cancer, cervical cancer



Slika 8. Matrica konfuzije random forest klasifikatora

Figure 8. Confusion matrix of random forest
From lefto to right/top to bottom:
breast cancer, colon cancer, liver cancer, lung cancer, cervical cancer

Diskusija

Implementacijom tri različita klasifikatora, izvršena je klasifikacija kancera na osnovu mutacija pojedinačnih baza u okviru čovekovog genoma. Cilj istraživanja bila je evaluacija performansi svakog od klasifikatora u zavisnosti od različitih parametara. Kao što je prikazano na slikama 4 i 5, najveću tačnost u klasifikaciji imala je logistička regresija; random forest je imao manju, a SVM najmanju tačnost. Logistička regresija se pokazala kao najbolja za klasifikaciju nad malom bazom podataka zbog njenih karakteristika. Ona novi uzorak upoređuje sa svakom klasom pojedinačno, i za svaku klasu računa verovatnoću da uzorak pripada toj klasi. Za razliku od logističke regresije, klasifikator maksimalne margine uzorak svrstava prema tome gde se on nalazi u odnosu na marginu koja deli klase. Sa malim brojem podataka margina ne može da razdvoji klase u potpunosti,

odnosno margina nije dovoljno jasno definisana, pa se veliki broj uzoraka koji pripada određenoj klasi nalazi sa suprotne strane margine. Zbog toga ovaj klasifikator daje najlošije rezultate.

Pri razvrstavanju mutacija u klase, za okolinu je uzeta po jedna baza sa obe strane mesta mutacije. Ako bi se za okolinu uzeo veći broj baza (po dve ili više), to bi povećalo broj klasa u koje se mutacije svrstavaju, i pretpostavlja se da bi se time postigle veće razlike između generisanih potpisa za svaki tip kancera. Za takvu analizu potrebna je znatno veća baza podataka nego što je baza nad kojom je rađeno istraživanje, što dovodi do činjenice da proširenje istraživanja u smeru povećanja okoline mutacija nije izvodljivo nad korišćenom bazom.

Zaključak

Iz dobijenih rezultata moguće je zaključiti da se povećanjem minimalnog broja supstitucija po pacijentu povećava i tačnost klasifikacije, čime se potvrđuje hipoteza sa početka istraživanja. Razlog toga je što podaci o supstitucijama pacijenata koje se prosleđuju klasifikatorima nose više informacija, te su i razlike u potpisima kod različitih tipova kancera tada više izražene.

Ukoliko posmatramo broj potpisa generisanih pomoću NNMF-a, postoji ograničenje koliko se najviše potpisa može generisati, i ono je jednako broju početnih klasa mutacija koji iznosi 96. Približavanjem broja generisanih potpisa broju početnih klasa mutacija, očekivano je da će se generisani potpisi poistovetiti sa klasama tipova mutacija, te NNMF nema uticaja na generisanje potpisa. S druge strane, ako se generiše broj potpisa koji teži jedinici, moguće je da se pojave različiti tipovi kancera sa istim dominantnim potpisima, što isto dovodi do velike greške pri klasifikaciji. Iz rezultata je moguće zaključiti da se data metoda analiziranja i klasifikacije tipa raka može koristiti nad sličnim vrstama podataka, sa pretpostavkom da se stepen uspešnosti klasifikacije može poboljšati proširivanjem same baze, odnosno dodavanjem još podataka novih pacijenata u bazu nad kojom se vrši istraživanje.

Literatura

ACS (The American Cancer Society medical and editorial content team) 2018. https://www.cancer.org/cancer/cancer-causes/genetics/family-cancer-syndromes.html#written_by

Alexandrov L. B., Nik-Zainal S., Wedge D. C., Aparicio S., Behjati S., Biankin A. V., et al. 2013. Signatures of mutational processes in human cancer. *Nature*, **500**: 415.

Ferlay J., Soerjomataram I., Dikshit R., Eser S., Mathers C., Rebelo M., et al. 2013. Cancer incidence and mortality worldwide: Sources, methodsand major patterns in GLOBOCAN 2012. *International Journal of Cancer*, **136**: E359.

Hanahan D., Weinberg R. A. 2000. The Hallmarks of Cancer Review. *Cell*, **100** (1): 57.

Nationaal cancer institute 2015. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer?redirect=true

Soh K. P., Szczurek E., Sakoparnig T., Beerenwinke N. 2017. Predicting cancer type from tumour DNA signatures. *Genome Medicine*, **9** (1): 104.

Diana Sekulić and Tijana Aleksić

Cancer Classification by Microarray Analysis

Establishing the cancer type and site of origin is important in determining the most appropriate course of treatment for cancer patients. Computer analysis and machine learning methods play a big role in modern medicine and cancer diagnosing. Recently there have been several studies that indicate that mutations in DNA can indicate the presence of cancer cells (Alexandrov *et al*. 2013).

Using patients' sequenced DNA, data on somatic single point mutations that occurred in the DNA of patients who had a specific type of cancer was collected. Five different types of cancer were used. Data on breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), liver cancer (LINC), lung squamous cell carcinoma (LUSC) and uterine corpus endometrial carcinoma (UCEC).

In this paper, the performances of three different classifiers were compared. Support vector machines (SVM), logistic regression and random forest algorithm were implemented and tested on the data. Figure 5 presents a graph that shows how the accuracy of each classifier is changing when the minimal number of substitutions per patient increases. As the minimal number of substitutions increases, the accuracy of each classifier increases as well. Figure 6 presents a graph that shows how the accuracy of each classifier is changing when the number of signatures that are generated in non-negative matrix factorization (NNMF) is increasing. Based on that, the overall classification evaluation was done with the data of patients who had more than 50 substitutions and 22 signatures generated in the factorization process. In the overall classification evaluation, logistic regression had the highest accuracy of 88.8% and precision of 90.7%. Random forest had the second best accuracy of 84.8% and precision of 84.7%, while support vector machines had the lowest accuracy of 76.8% and precision of 79.0%. Confusion matrices of each classifier are presented in figures 7-9.