Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.39
DOI: https://doi.org/10.24867/03BE04Jankovic

EVALUACIJA PERFORMANSI KONSENZUS KLASTEROVANJA NAD HISTOPATOLOŠKIM SLIKAMA TUMORA DOJKE

EVALUATION OF CONSENSUS CLUSTERING PERFORMANCE ON HISTOPATHOLOGICAL BREAST CANCER IMAGES

Milica Janković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – Energetika, elektronika i računarstvo

Kratak sadržaj – *U ovom radu prikazana je analiza i izdvajanje obeležja sa histopatoloških slika tumora dojke kako bi se postiglo njihovo klasterovanje na benigne i maligne uzorke. Korišćeno je šest metoda za izdvajanje obeležja, PCA metoda redukcije dimenzionalnosti, konsenzus i polunadgledano klasterovanje i adjusted rand indeks kao mera validacije klasterovanja.*

Ključne reči: *Konsenzus klasterovanje, polunadgledano učenje, histopatološke slike, tumor dojke, pca*

Abstract – *In this paper we present histopathological breast cancer image analysis, feature extraction, and their clustering into benign and malignant. We used six methods for feature extraction, PCA for dimensionality reduction, ensamble clustering and semi-supervised learning were evaluated and adjusted rand index was used as an external validation measure.*

Keywords: *Consensus clustering, semi-supervised learning, histopathological images, breast cancer, PCA*

1. UVOD

Prema istraživanjima Svetske zdravstvene organizacije (*World Health Organization*), u 2012.godini je umrlo 8.2 miliona ljudi od kancera i očekuje se da će se ta brojka povećati na 27 miliona do 2030.godine. Kancer dojke se najčešće javlja kod žena, a od raka dojke mogu oboleti i muškarci. Rak dojke je stotinu puta češći kod žena, nego kod muškaraca. Zloćudni ili maligni tumori dojke su drugi po redu uzrok smrti, odmah posle karcinoma pluća. Od svih smrtnih ishoda malignih tumora, smrtnost od karcinoma dojke uzima oko 20% [1].

Biopsija je jedini način da se sa sigurnošću utvrdi priroda uočenih promena. Nakon biopsije vrši se histopatološka analiza uzoraka tkiva i klasifikacija benignih i malignih promena od strane patologa. Manuelna analiza i klasifikacija histopatoloških slika donekle je subjektivan proces, dodatno podložan greškama usled umora i obima posla. Računarska podrška i automatska klasifikacija histopatoloških slika bile bi značajan doprinos u svakodnevnom radu patologa, kao podrška proceni patologa donešenoj na osnovu znanja i iskustva.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila prof. dr Tatjana Lončar Turukalo.

Kako bi se poboljšala histopatološka analiza uloženo je dosta truda, posebno kako bi se poboljšala automatska klasifikacija malignih i benignih slika za kompjutersku dijagnostiku. Na osnovu svih dosadašnjih radova se može zaključiti da je glavna prepreka, u razvoju novih metoda za histopatološku analizu nedostatak dovoljno velike anotirane baze podataka koja je dostupna široj javnosti za evaluaciju algoritama koji bi omogućili finu klasifikaciju uzoraka, ne samo na benigne i maligne, već i utvrđivanje podvrsta tumora.
Stoga je u radu [2] korišćena i predstavljena baza podataka *BreaKHist*, sakupljena upravo da omogući olakšan razvoj algoritama za klasifikaciju ovih slika. U tom radu je prvo vršeno izdvajanje obeležja, a zatim klasifikacija gde je postignuta tačnost od 80% do 85% ovisno o faktoru uvećanja slike.

U ovom radu je korišćena ista baza podatka i isti skup obeležja, ali je umesto klasifikacije vršeno klasterovanje. Evaluiraju se performanse jedne metode za nenadgledano učenje i jedne metode za polunadgledano učenje pri grupisanju histopatoloških slika na osnovu teksturalnih obeležja.

2. MATERIJALI I METODE

2.1. Baza podataka

U ovom radu korišćena je pomenuta *BreaKHist* baza podataka, koja sadrži slike koje prikazuju mikroskopski prikaz biopsije benignih i malignih tumora dojke.

Digitalne slike su dobijene tako što je mikroskop povezan sa digitalnim fotoaparatom. One su 24-bitne slike u boji (8 bita za svaki RGB kanal) sa uvećanjima 40×, 100×, 200× i 400×. Veličina slike je 700 × 460 piksela, bez kompresije i normalizacije. Baza slika sadrži 7909 slika podeljenih na benigne i maligne tumore.

Uvećanje	Benigni	Maligni	Ukupno
40x	625	1370	**1995**
100x	644	1437	**2081**
200x	623	1390	**2013**
400x	588	1232	**1820**
Ukupno	**2480**	**5429**	**7909**
Pacijenti	24	58	82

Tabela 1. Raspodela slika po klasama i uvećanjima

2.2. Metode

2.2.1. Izdvajanje obeležja

Pre izdvajanja obeležja vršena je predobrada slika normalizacijom. Algoritam koji se koristi automatski pronalazi ispravan vektor bojenja i vrši korekciju boje [3]. Nakon korekcije vektora bojenja, vrši se korekcija inteziteta.

Za izdvajanje obeležja korišćeni su uglavnom teksturalni deskriptori koji spadaju u najčešće korišćene teksturalne deskriptore. Izbor ove vrste obeležja ima smisla na manjim uvećanjima, dok na uvećanju od 400x pored ovih atributa mogla je biti vršena i segmentacija i karakterizacija ćelija, ali je zbog obima izostavljeno iz ovog istraživanja.

LBP **(***Local Binary Patterns*) je operator koji izračunava raspodelu binarnih oblika u kružnom susedstvu svakog piksela [4]. Susedstvo karakteriše poluprečnik R i broj suseda P. Intenzitet susednih piksela se poredi sa intenzitetom centralnog piksela: svakom od P suseda dodeljuje se vrednost 1 ako je intenzitet posmatranog piksela veći ili jednaki intenzitetu centralnog piksela, u suprotnom dodeljuje se vrednost 0. Za svaki piksel se na ovaj način formira binarni obraza. LBP se za svaki piksel u kompaktnom zapisu može prikazati kao binarni kod:

$$LBP(p) = \sum_{i=0}^{P-1} 2^i \delta(f(q_i) - f(p)) \qquad (1)$$

gde su $f(q_i)$ i $f(p)$ intenziteti piksela q_i i p, a δ je *Kronecker* funkcija. Ovde se koristi rotaciono-invarijantni uniformni LBP, a histogram LBP koda se koristi kao teksturalni deskriptor.

CLBP (*Completed Local Binary Pattern*) je jedna od najnovijih varijanti LBP-a [5]. Ima tri komponente izvučene iz lokalnog regiona: centralni piksel, znak i intenzitet. Centralnom pikselu je dodeljen binarni kod nakon globalne primene praga. Za komponente znaka i intenziteta, razmatrano je susedstvo centralnog piksela radijusa R veličine P suseda, slično LBP-u. Tri komponete se spajaju kako bi se formirao završni CLBP histogram. Ovde se koristi rotaciono-invarijantni uniformni CLBP sa parametrima P = 24 i R = 5.

LPQ (*Local Phase Quantization*) predstavlja novu metodu koja je otporna na zamućenje u slikama. Ona se bazira na kvantizaciji DFT faze u lokalnim prozorima slike [6]. LPQ operator se računa lokalno za svaki piksel, a rezultati se predstavljaju pomoću histograma. U ovom radu je korišćena varijacija LPQ-a koja se zove LPQ-TOP, koja podrazumeva korišćenje različitih vrednosti parametara.

GLCM (*Gray-Level Co-Occurrence Matrices*) se koristi za karakterizaciju teksture slike i bazira se na matricama koje ukazuju na prostornu zavisnost intenziteta. U ovom radu, za izračunavanje GLCM-a se koriste četiri pravca sa uglovima od 0°, 45°, 90°, 135° i osam nivoa intenziteta. Zatim se nad datim GLCM matricama izračunavaju *Haralick*-ovi atributi [7]. Završni vektor se formira izračunavanjem i konkatenacijom prosečnih vrednosti 13-dimenzionalnog vektora za sva četiri smera.

PFTAS (*Parameter-Free Threshold Adjacency Statistics*) je verzija TAS bez parametara [8]. Prvi korak je binarizacija slike upotrebom tri opsega: $[\mu + \sigma , \mu - \sigma]$, $[\mu - \sigma , 255]$ i $[\mu , 255]$, gde je σ standardna devijacija piksela. Prag μ se izračunava na osnovu Otsu metode. Nakon toga se prebrojavaju beli pikseli koji imaju *i* (0-8) belih suseda i formira odgovarajući histogram. Završni vektor obeležja se dobija tako što se 81-dimenzionalni vektor obeležja poveže sa svojom bitski negativnom verzijom.

ORB (*Oriented FAST and Rotated BRIEF*) se zasniva na FAST detektoru ključnih tačaka i BRIEF deskriptoru ključnih tačaka [9]. U ovom radu korišćena je OpenCV implementacija algoritma, bez promene parametra. Najbolji rezultati se postižu sa 500 ključnih tačaka.

Nakon izdvajanja obeležja vrši se redukcija dimenzionalnosti primenom analize osnovnih komponenti (PCA). Cilj PCA je da predstavi skup uzoraka što tačnije u prostoru sa manjim brojem dimenzija. Očuvanje informacija postiže se očuvanjem varijanse podataka u što većoj meri, za svaki skup je određen prostor sa manjim brojem dimenzija, tako da se u tom prostoru sačuva više od 99% varijanse iz prvobitnog prostora obeležja.

2.2.2. Klasterovanje

Klasterovanje je tehnika istraživanja podataka koja objekte (koji se opisuju atributima) sličnih osobina deli u grupe (klastere), čineći ih preglednijim i korisnijim.

Klasterovanje na osnovu akumulacije dokaza (EAC - *Evidence Accumulation Clustering*) je pristup koji koristi veći broj particija dobijenih nekim jednostavnijim algoritmom klasterovanja da bi formirao novu matricu sličnosti koja reflektuje sličnost uzoraka na osnovu verovatnoće njihovog pojavljivanja u istom klasteru procenjene nad ansamblom particija [10]. Prvo se počinje višestrukom podelom podataka korišćenjem brzog *k-means* (KM) u veliki broj kompaktnih i malih klastera. Slučajnom inicijalizacijom algoritma KM se dobijaju različite particije, a i broj klastera pri svakom pokretanju algoritma bira se na slučaj iz opsega $\left[\frac{\sqrt{N}}{2}, \sqrt{N}\right]$, gde je N ukupan broj raspoloživih slika. Nakon toga se podaci iz particija mapiraju u matricu koasocijacije koja predstavlja novu meru sličnosti između slika:

$$co\_assoc(i,j) = votes_{ij}/M \qquad (2)$$

gde je M broj particija u ansamblu, a $votes_{ij}$ je broj koliko puta je par uzoraka *(i, j)* dodeljen istom klasteru u M klasterovanja. Zatim je nad matricom koasocijacije vršeno hijerarhijsko klasterovanje *average linkage* metodom, a kao parametar korišćen je samo broj klastera.

Polunadgledano konsenzus klasterovanje (SSCC - *Semi-Supervised Consensus Clustering*) se bavi problemom pronalaženja konsenzusne particije podataka koristeći prilikom pronalaženja rešenja znanje o labelama određenog procenta raspoloživih slika [11]. Ako sa $\mathcal{L} \subset \mathcal{I}$ označimo indekse podataka koji su labelirani i sa $l_i \in \{1, \dots, k\}$ labelu koja odgovara *i*-toj slici, $i \in \mathcal{L}$.

Problem koji rešava SSCC može da se formuliše kao problem minimizacije sa ograničenjem:

$$\phi^* \in \arg \min_{\hat{\phi} \in \mathcal{I} \to \{1,\dots,k\}} \sum_{u=1}^{m} d(\hat{\phi}, \phi_u) \quad (3)$$

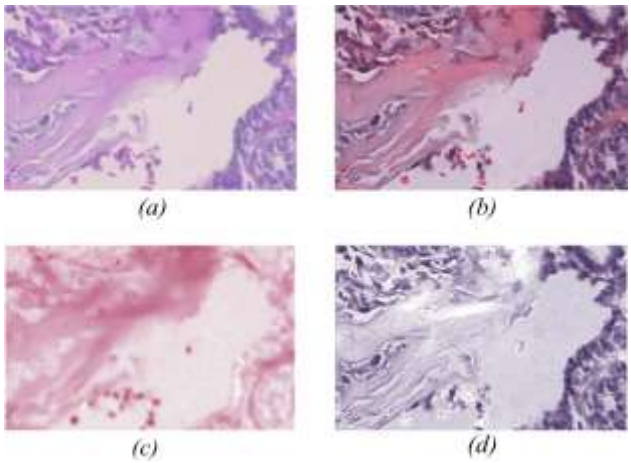
gde je $\hat{\phi}(i) = l_i$ za svaki $i \in \mathcal{L}$, a $\phi_u \in \mathcal{J}_u \to \{1, \dots, k_u\}$ je funkcija koja koduje podelu podskupa tačaka podataka indeksovanih sa $\mathcal{J}_u \subseteq \mathcal{I} = \{1, \dots, N\}$ u k_u klastera. U ovom radu ansambl je konstruisan korišćenjem ansambla od 200 particija KM algoritma i 20% poznatih labela.

Mere validacije klasterovanja je vršena primenom *adjusted rand* indeks (ARI) kao eksterne mere za validaciju klasterovanja.

3. REZULTATI

3.1. Rezultati normalizacije

Normalizacija boje je vršena nad slikama iz baze podatka za sva tri RGB kanala. Jedan primer efekata ove normalizacije prikazan je na Slici 1. Na ovaj način izbegnute su varijacije u obeležjima usled neujednačene obojenosti uzoraka koja zavisi i od kvaliteta boja i od starosti uzorka.



Slika 1. (a) Prikaz malignog tumora; (b) Prikaz slike nakon izdvajanja boje (hematoksilin); (c) Prikaz slike nakon izdvajanja boje (eozina) i (d) Prikaz slike nakon normalizacije

3.2. Rezultati izdvajanja obeležja

Nakon normalizacije vršeno je izdvajanje obeležja. Za svako uvećanje i za svaku grupu obeležja izdvajanje je vršeno nezavisno, a dobijene su matrice dimezije *N x d* gde je *N* broj vrsta koji odgovara broju raspoloživih slika za dato uvećanje (Tabela 1), a *d* broj kolona koji odgovara broju obeležja za sva tri RGB kanala prikazan u Tabeli 2.

Naziv	LBP	CLBP	LPQ	GLCM	PFTAS	ORB
Broj obeležja	30	4056	768	39	162	96

Tabela 2. Ukupan broj obeležja iz svake grupe

3.3. Rezultati redukcije dimenzionalnosti

Nad matricama obeležja vrši se redukcija dimenzionalnosti PCA metodom tako da se očuva 99% varijanse. U Tabeli 3 prikazan je novi broj obeležja nakon redukcije dimenzionalnosti. Potrebno je uočiti da je za različita uvećanja PCA analiza rezultovala različitim brojem obeležja.

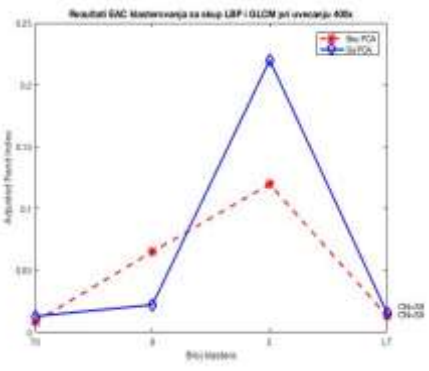
Uvećanje	LBP	CLBP	LPQ	GLCM	PFTAS	ORB
40x	11	1304	140	8	31	75
100x	11	1237	144	9	32	74
200x	12	1191	147	9	30	66
400x	12	1205	166	9	29	65

Tabela 3. Vrednosti kolona matrice obeležja nakon PCA

3.4. Rezultati klasterovanja

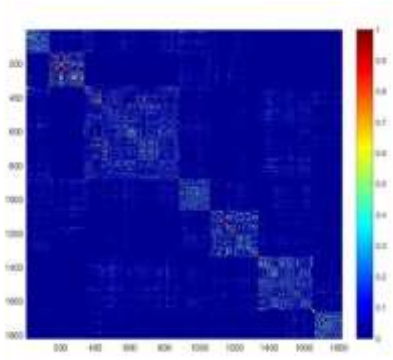
Kod EAC klasterovanja za broj klastera se uzimaju vrednosti 2, 8 (ukupan broj podvrsta tumora), 70 (radi procene načina grupisanja uzoraka) i broj dobijen *life-time* (LT) kriterijumom, odnosno broj klastera koji se dobija u najdužem opsegu presecanja dendrograma. Dok se kod SSCC klasterovanja za broj klastera zadaju vrednosti 2, 8, 50 i 70. Klasterovanje je vršeno za svaku grupu obeležja zasebno za različit broj klastera sa i bez PCA. Nakon toga za skup svih obeležja zajedno za različit broj klastera sa i bez PCA. I na kraju za skup obležja sa i bez PCA, koji je formiran spajanjem onih obeležja koja su postigla najveću tačnost u diplomskom radu [12] u kojem je vršena klasifikacija nad istim skupom, takođe za različit broj klastera. Ovde će, zbog obimnosti, biti prikazani samo neki od rezultata.

Rezultati EAC klasterovanja sa i bez PCA za skup koji je dobijen spajanjem LBP i GLCM obeležja za različit broj klastera (2, 8, 70, *LT* kriterijum) za uvećanje 400x su prikazani na Slici 2. Zapaža se da se sa smanjenjem broja klastera poboljšavaju rezultati dok primena PCA ovde dobro utiče posebno kada je broj klastera dva gde je ARI 0.22.



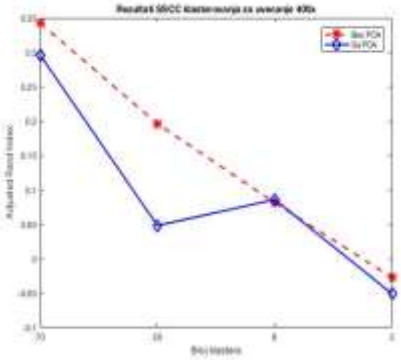
Slika 2. Rezultati EAC klasterovanja sa i bez PCA za skup LBP i GLCM obeležja za različit broj klastera

Na Slici 3. prikazana je sortirana matrica koasocijacije za slučaj osam klastera gde se uočava slaba sličnost među uzorcima iz istog klastera, koja se ogleda i u malom ARI indeksu (Slika 2).



Slika 3. Sortirana matrica koasocijacije nad LBP i GLCM skupom obeležja bez PCA (k=8)

Na Slici 4. prikazani su rezultati SSCC klasterovanja za skup svih obeležja združeno pre i nakon primene PCA za različit broj klastera (2, 8, 50, 70). Na osnovu slike uočava se malo poboljšanje kada je broj klastera 70 gde je ARI 0.3426, a to je najbolji rezultat postignut u odnosu na druga uvećanja. Takođe se uočava da se rezultati pogoršavaju sa smanjivanjem broja klastera i primenom PCA, uglavnom jer se manje grupe uzoraka izoluju u menje klastere, dok je većina uzoraka u jednom klasteru.



Slika 4. Rezultati SSCC klasterovanja sa i bez PCA za sva obeležja i različit broj klastera (2, 8, 50, 70)

Za uvećanje 400x u Tabeli 4. prikazana je raspodela uzoraka po klasterima, za broj klastera 8. U vrstama su naznačene originalne labele. Uočava se veliko mešanje među uzorcima što ukazuje na veliku sličnost između nekih benignih i malignih tumora i potrebu za dodatnim obeležjima ili njihovom nelinearnom transformacijom u cilju bolje separabilnosti uzoraka.

	Broj klastera							
	1	2	3	4	5	6	7	8
Benigni	92	118	0	87	121	40	86	44
Maligni	187	0	246	200	52	222	20	305

Tabela 4. Rezultati SSCC bez PCA za k=8

4. ZAKLJUČAK

Klasterovanjem, čak i primenom polunadgledanog učenja nije dobijeno zadovoljavajuće grupisanje uzoraka po klasama. Uočeno je veliko izdvajanje određenih grupa uzoraka, što u postupku hijerarhijskog klasterovanja uslovljava da se ti specifični uzorci prvi grupišu u manje klastere, dok veliki broj benignih i malignih uzoraka ostaje u istom klasteru čak i pri povećanju broja klastera na 50 ili 70. Kako bi se poboljšali rezultati klasterovanja jedan od mogućih pristupa bi bio izbor različitih parametara prilikom izdvajanja obeležja i testiranje njihovog uticaja kako na EAC tako i na SSCC. Pored toga potrebno je razmotriti i druga obeležja, posebno za veća uvećanja. Dodatni izazov bi bio pokušaj klasterovanja uzoraka u podtipove malignih i benignih tumora, kao i pokušaj klasterovanja po pacijentima.

5. LITERATURA

[1] World Health Organization: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] F. A. Spanhol *et al*., "A Dataset for Breast Cancer Histopathological Image Classification", IEEE Transactions on Biomedical Engineering, str. 1455-1463, 2016.

[3] M. Macenko *et al*, "A Methode For Normalizing Histology Slides For Quantitive Analysis", IEEE International Symposium on Biomedical Imaging, str. 209., 2013.

[4] T. Ojala *et a*l., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Trans. Pattern Anal.Mach. Intell., str. 971–987, 2002.

[5] Z. Guo *et al*.,, "A completed modeling of local binary pattern operator for texture classification", IEEE Trans. Image Process, str. 1657–1663, 2010.

[6] V. Heikkilin *et al*, "Blur insensitive texture classification using local phase quantization", Proc. 3rd Int. Conf. Image Signal Process., str. 236–243, 2008.

[7] R. Haralick *et al*, "Textural features for image classification", IEEE Trans. Syst. Man Cybern., str. 610-621, 1973.

[8] L. P. Coelho *et al*, "Structured literature image finder: extracting information from text and images in biomedical literature", New York : Springer, str.121, 2010.

[9] E. Rublee *et al*, "ORB: An efficient alternative to SIFT or SURF", Proc. IEEE Int. Conf. Comput. Vision, str. 2564–2571, 2011.

[10] L.N. Fred, "Data clustering using evidence accumulation", IEEE Object recognition supported by user interaction for service robots., 2002.

[11] H. Aidos, "Semi-Supervised Consensus Clustering for ECG Pathology Classification", Proc European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases, str. 150-164, 2015.

[12] A. Antić, "Klasifikacija histopatoloških slika tumora dojke", Diplomski rad, Novi Sad : FTN, 2018.

Kratka biografija:

Milica Janković rođen je u Doboju 1991. god. Master rad na Fakultetu tehničkih nauka iz oblasti Energetike, elektronike i telekomunikacija – Obrada signala, odbranila je 2018.god.
Kontakt: milica24111991@live.com