Evaluacija algoritma k-srednjih vrednosti pri klasterizaciji genskih ekspresija karcinoma

Studentski rad

Katarina Ćoćić, Boris Knežević
student master studija
Fakultet tehničkih nauka
Novi Sad, Srbija
katarinacocic95@gmail.com
borisknezevic@uns.ac.rs

Sažetak—**Algoritmi klasterizacije spadaju u tehnike nenadgledanog učenja koje pokušavaju da utvrde strukturu podataka. Ove metode se sve češće koriste u cilju grupisanja genskih ekspresija, formiranja odgovarajuće terapije, kao i otkrivanja novih podtipova raka. Međutim, veliki broj obeležja otežava proces klasterizacije. U ovom radu vršena je evaluacija performansi klasterizacije na osnovu algoritma *k-srednjih vrednosti*. Analiza je izvršena nad normalizovanim i nenormalizovanim podacima sa dve vrste rastojanja (euklidsko rastojanje i korelacija). Za validaciju celog postupka su iskorišćene interne (Silhouette i Calinski-Harabasz) i eksterne mere validacije (ARI - Adjusted Rand Index). Na osnovu rezultata se može zaključiti da nad različitim skupovima podataka algoritam k-srednjih vrednosti pokazuje drugačije performanse, pri čemu postoji nedostatak u robusnosti rešenja, kao i osetljivost na izbor parametara.**

***Ključne reči* – k–srednjih vrednosti; validacija; genske ekspresije; kancer**

I. UVOD

Karcinom je jedna od bolesti u svetu koja godišnje odnese značajan broj života [1]. Ćelije raka se ne ponašaju kao i ostale ćelije organizma, već nastaju usled somatskih promena u strukturi genoma na nivou DNK. Osim toga, odlikuju se drugačijom strukturom, brže se dele, pa je samim tim otežana njihova analiza [2]. Iz tog razloga, čak i male promene u ekspresiji gena mogu biti od velikog značaja za pronalaženje efikasnije terapije i rane djagnostike [3].

Metode klasterizacije se danas neretko upotrebljavaju za klasterizaciju genskih ekspresija. Baziraju se na grupisanju uzoraka sa sličnim ekspresijama gena, ili na grupisanju gena koji imaju slične ekspresije u uzoračkom skupu. Uprkos velikom broju novih klasterizacionih tehnika boljih performansi, u većini radova klasterizacija genskh ekspresija vrši se algoritmom hijerarhijske klasterizacije, veoma intuitivnim i jednostavnim za interpretaciju. [4]

Ovaj rad analizira 35 javno dostupnih skupova podataka, koji su dobijeni pomoću dve vrste mikročipa – Affymetrix i cDNA [5]. Mali broj uzoraka, velika dimenzionalnost i šum svojstven načinu akvizicije podataka, čine podatke zahtevnim za analizu.

II. METODE

A. Skupovi podataka

Algoritmi klasterizacije i validacije su izvršeni nad trideset pet skupova podataka [5]. Postoji nekoliko ključnih razlika između datih skupova: čip uz pomoć kog su podaci prikupljeni, vrsta tkiva koje odgovarajući skup opisuje, broj uzoraka u skupu *N*, broj klasa *k*, raspodela uzoraka po klasama, dimenzionalnost skupa (broj analiziranih genskih ekspresija) pre i posle redukcije dimenzionalnosti (*m* i *d*, respektivno). Analiza je izvršena samo na dostupnim skupovima sa redukovanim brojem obeležja. Podaci su dobijeni korišćenjem dve mikročip tehnologije: cDNA i Affymetrix. cDNA skupova ima četrnaest, dok su Affymetrix čipovi korišćeni kod dvadeset jednog skupa podataka (Tabela 1).

Podaci dobijeni pomoću Affymetrix čipa se odlikuju velikom varijabilnošću i nalaze se u opsegu od 10 do 16 000. Snimanje se zasniva na broju kopija ribonukleinske kiseline, koje su pronađene u samoj ćeliji [6].

Razlika cDNA u odnosu na Affymetrix čip je u tehnologiji izrade i načinu snimanja. Dobijene vrednosti predstavljaju odnos broja ribonukleinskih kopija u odnosu na broj u kontrolnoj ćeliji [6]. cDNA podaci se odlikuju manjom varijabilnošću i imaju manji dinamički opseg.

Raspoloživi skupovi podataka analizirani su bez normalizacije i korišćenjem z-normalizacije. Z–normalizovan skup podataka okarakterisan je obeležjima nulte srednje vrednosti i jedinične standardne devijacije. Potrebno je napomeuti da je ova vrsta normalizacije osetljiva na postojanje artefakata u podacima.

Pre primene određene metode klasterizacije, potrebno je odrediti meru sličnosti među uzorcima. Kao mere sličnosti korišćene su euklidsko rastojanje između uzoraka i korelacija.

B. Klasterizacija

Kao algoritam klasterizacije odabran je algoritam k-srednjih vrednosti [7], [8]. To je jedan od najšešće korišćenih klasterizacionih algoritama. Veoma je intuitivan, računarski brz i jednostavan za implementaciju. Klasteri dobijeni ovim algoritmom su predstavljeni preko svojih centara koji opisuju sve uzorke unutar odgovarajućih klastera.

Na početku izvršavanja algoritma, potrebno je definisati broj klastera i inicijalizovati te klastere proizvoljnim postavljanjem centroida klastera. Inicijalni centroidi se postavljaju tako što se nasumično bira *k* uzoraka iz skupa podataka, gde *k* predstavlja broj klasa u odgovarajućem skupu (Tabela 1). Uzorci se smeštaju u onaj klaster čiji centroid im je najbliži. Položaj centroida se menja u iterativnom postupku tako da se minimizuje suma kvadrata rastojanja unutar klastera.

TABELA 1. RASPOLOŽIVI SKUPOVI PODATAKA GENSKIH EKSPRESIJA KARCINOMA

Tkivo	Skup podataka	Čip	*N*	*k*	Raspodela uzoraka	*m*	*d*
Krv	Armstrong-V1	Affymetrix	72	2	28, 48	12582	1081
Krv	Armstrong-V2	Affymetrix	72	3	24, 20, 28	12582	1081
Pluća	Bhattacharjee	Affymetrix	203	5	139, 17, 6, 21, 20	12600	1543
Dojka, debelo crevo	Chowdary	Affymetrix	104	2	62, 42	22283	182
Bešika	Dyrskjot	Affymetrix	40	3	9, 20, 11	7129	1023
Koštana srž	Golub-V1	Affymetrix	72	2	47, 25	7129	1877
Koštana srž	Golub-V2	Affymetrix	72	3	38, 9, 25	7129	1877
Pluća	Gordon	Affymetrix	181	2	31, 150	12533	1626
Debelo crevo	Laiho	Affymetrix	37	2	8, 29	22883	2202
Mozak	Nutt-V1	Affymetrix	50	4	14, 7, 14, 15	12625	1377
Mozak	Nutt-V2	Affymetrix	28	2	14, 14	12625	1070
Mozak	Nutt-V3	Affymetrix	22	2	7, 15	12625	1152
Mozak	Pomeroy-V1	Affymetrix	34	2	25, 9	7129	857
Mozak	Pomeroy-V2	Affymetrix	42	5	10, 10, 10, 4, 8	7129	1379
Više različitih tipova tkiva	Ramaswamy	Affymetrix	190	14	11, 10, 11, 11, 22, 10, 11, 10, 30, 11, 11, 11, 11, 20	16063	1363
Krv	Shipp	Affymetrix	77	2	58, 19	7129	798
Prostata	Singh	Affymetrix	102	2	58, 19	12600	339
Više različitih tipova tkiva	Su	Affymetrix	174	10	26, 8, 26, 23, 12, 11, 7, 27, 6, 28	12533	1571
Dojka	West	Affymetrix	49	2	25, 24	7129	1198
Koštana srž	Yeoh-V1	Affymetrix	248	2	43, 205	12625	2526
Koštana srž	Yeoh-V2	Affymetrix	248	6	15, 27, 64, 20, 79, 43	12625	2526
Krv	Alizadeh-V1	cDNA	42	2	21, 21	4022	1095
Krv	Alizadeh-V2	cDNA	62	3	42, 9, 11	4022	2093
Krv	Alizadeh-V3	cDNA	62	4	21, 21, 9, 11	4022	2093
Koža	Bittner	cDNA	38	2	19, 19	8067	2201
Mozak	Bredel	cDNA	50	3	31, 14, 5	41472	1739
Jetra	Chen	cDNA	180	2	104, 76	22699	85
Pluća	Garber	cDNA	66	4	17, 40, 4, 5	24192	4533
Više različitih tipova tkiva	Khan	cDNA	83	4	29, 11, 18, 25	6567	1069
Prostata	Lapointe-V1	cDNA	69	3	11, 39, 19	42640	1625
Prostata	Lapointe-V2	cDNA	110	4	11, 39, 19, 41	42640	2496
Mozak	Liang	cDNA	37	3	28, 6, 3	24192	1411
Endometrijum	Risinger	cDNA	42	4	13, 3, 19, 7	8872	1771
Prostata	Tomlins-V1	cDNA	104	5	27, 20, 32, 13, 12	20000	2315
Prostata	Tomlins-V2	cDNA	92	4	27, 20, 32, 13	20000	2315

*N – broj uzoraka, k – broj klastera, m – broj obeležja pre smanjivanja dimenzionalnosti, d – broj obeležja nakon smanjivanja dimenzionalnosti

U ovom radu, algoritam k-srednjih vrednosti je ponovljen dvadeset i pet puta na svakom skupu podataka sa nasumičnom inicijalizacijom centara klastera postavljanim na raspoložive uzorke i brojem klasa postavljenim na stvaran broj klasa.

C. Validacija

Za procenu kvaliteta rezultata dobijenih primenom algoritma k-srednjih vrednosti korišćene su eksterne mere validacije, koje koriste poznate referentne vrednosti i interne mere validacije koje analiziraju podatke iz istog klastera.

Od eksternih mera validacije, korišćen je ARI (*Adjusted Rand Index*) [9]. ARI poredi labele dobijene klasterizacijom sa originalnim labelama uzoraka uz korekciju za slučajna podudaranja. Ovaj indeks ukazuje na to koliko se dobro dobijeni klasteri poklapaju sa stvarnim klasama. Vrednost jedan se dostiže kada su originale labele i labele dobijene klasterizacijom identične.

Kao interna mera validacije, korišćena su dva indeksa: Silhouette (SIL) i Calinski-Harabasz (CH). *SIL* [10] indeks opisuje koliko je uzorak sličan uzorcima u klasteru u koji je svrstan, u poređenju sa uzorcima iz ostalih klastera. Izračunava se po sledećoj fomuli [10]:

$$\mathrm{s(i)} = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (1)$$

gde a(*i*) predstavlja rastojanje uzorka *i* od svih uzoraka iz istog klastera, a b(*i*) predstavlja rastojanje od uzoraka drugog klastera. Opseg vrednosti ovog indeksa je od -1 do +1, gde visoka vrednost indeksa ukazuje na to da je uzorak smešten u odgovarajući klaster. Ukoliko je ova vrednost niska ili negativna, to može da ukaze i na neadekvatan broj klastera.

Calinski- Harabasz indeks [11] koristi pretpostavku da dobra klasterizacija ima karakteristiku male sume kvadrata rastojanja u okviru klastera, odnosno velike sume kvadrata rastojanja između uzoraka iz različitih klastera. Što je vrednost ovog indeksa veća, to je klasterizacija kompaktnija. Izračunava se po sledećoj formuli [11]:

$$CH_k = \frac{SSb}{SSw} \times \frac{N-k}{k-1} \qquad (2)$$

gde je SSb međuklasna varijansa, SSw je unutarklasna varijansa, N je broj uzoraka, a k je broj klastera.

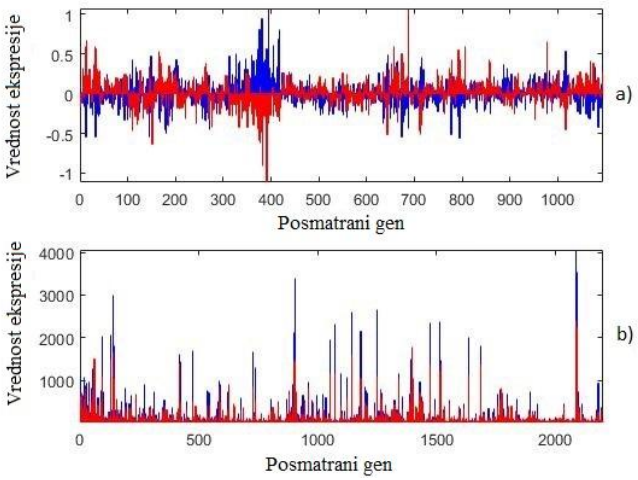
III. GLAVNI REZULTATI

A. Statistička analiza obeležja

Statistička analiza obeležja, ekspresija gena, kod svih skupova podataka vršena je sa ciljem utvrđivanja značaja pojedinih ekspresija na diskriminaciju podtipova karcinoma. Karakteristika obeležja koja je ispitivana jeste medijan obeležja, zavisno od klase u kojoj se uzorci nalaze. Medijani obeležja prikazani su barplot-ovima, gde je svaka klasa predstavljena različitom bojom.

Slika 1 prikazuje ilustrativni primer dva skupa obeležja, jednog dobijenog cDNA mikročipom, a drugog Affymetrix čipom.

Slika 1a prikazuje skup podataka Alizadeh-V1 prikupljen uz pomoć *cDNA* mikročipa i opisuje ekspresije gena kod pacijenata koji su oboleli od leukemije. Pacijenti su podeljeni u 2 klase. Na Slici. 1. se primećuje da određena obeležja imaju veoma različite medijane u zavisnosti od posmatrane klase. Pojedina obeležja sa pozitivnim medijanom u prvoj klasi imaju negativan medijan u drugoj klasi i obrnuto. Takva obeležja doprinose većoj diskriminativnosti između klasa i boljoj klasterizaciji uzoraka (Slika. 1).



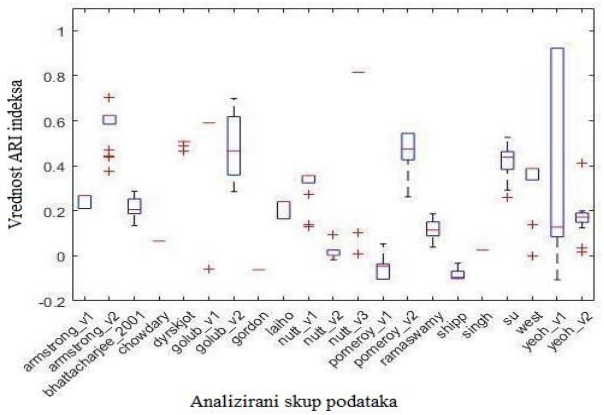
Slika 1. Medijan obeležja (vrednosti genske ekspresije) u skupu a) Alizadeh-V1 i b) Laiho

Slika 1b prikazuje skup podataka Laiho prikupljen uz pomoć *Affymetrix* mikročipa i opisuje ekspresije gena kod pacijenata koji boluju od karcinoma debelog creva. Pacijenti su, takođe, podeljeni u dve klase. Međutim, sada je preklapanje između medijana u različitim klasama mnogo veće, dok je broj diskriminativnih obeležja mnogo manji. U ovakvim slučajevima bi se mogla odraditi analiza obeležja, gde bi se diskriminativna obeležja sačuvala, a obeležja sa sličnim medijanima bi se smatrala šumom i odbacivala (Slika. 1). U ovom radu nije vršena redukcija dimenzionalnosti, već samo analiza atributa koja je ukazala na specifičnosti mikročip tehnologija i potrebu za korišćenjem različitih pristupa normalizaciji i merenju sličnosti između uzoraka za ove dve tehnologije.

B. Rezultati validacije

Performanse algoritma k-srednjih vrednosti značajno su zavisile od osobina skupa podataka, pre svega vrste mikročip tehnologije. Stoga su skupovi dve različite tehnologije zasebno razmatrani. Rezultati indeksa validacije su prikazani pomoću boxplot-ova koji sumiraju vrednosti u 25 realizacija nad svakim skupom podataka zasebno.

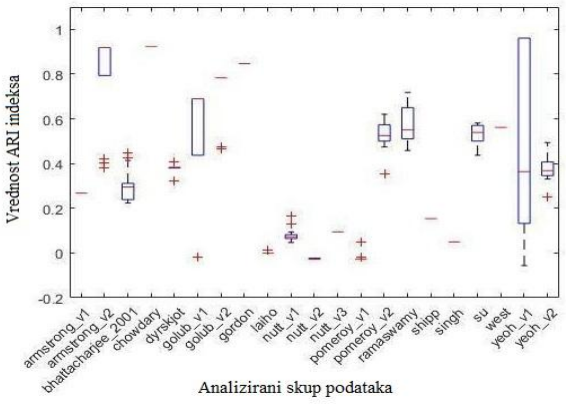
Baza dobijena pomoću Affymetrix čipa sadrži 21 skup podataka. Primenjena klasterizacija nad nenormalizovanim podacima sa euklidskim rastojanjem ne daje konzistentne rezultate za sve skupove podataka (Slika. 2). Varijabilnost algoritma k-srednjih vrednosti usled osetljivosti na incijalne uslove, predstavlja jedan od značajnijih problema koji utiče na robusnost dobijenih rešenja. Vrednost ARI indeksa je manja od 0,5 nad 18 skupova podataka.



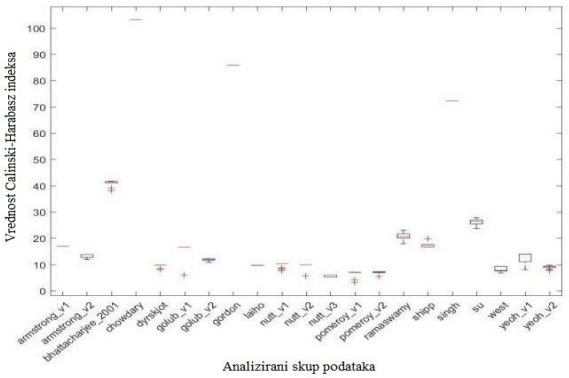
Slika 2. Vrednost ARI za nenormalizovane *Affymetrix* podatke, primenom algoritma k-srednjih vrednosti sa euklidskim rastojanje

Primenom korelacije umesto euklidskog rastojanja, uočava se da nekim skupovima korelacija, kao mera rastojanja, više pogoduje, pri čemu 10 skupova ima ARI iznad pomenute granice od 0,5. Vrednosti indeksa, u ovom slučaju, imaju manju varijabilnost što ukazuje na stabilniju klasterizaciju (Slika. 3.).

Calinski–Harabasz indeks pokazuje veoma male vrednosti za većinu skupova podataka (Slika. 4). *Silhouette* indeks znatno više varira i pokazuje najbolje performanse na originalnom skupu podataka sa euklidskim rastojanjem.

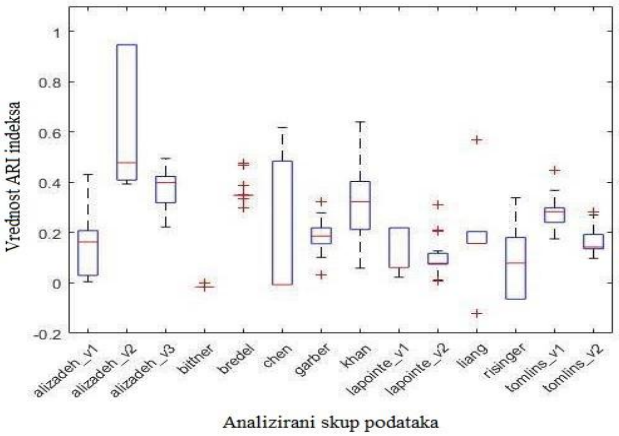


Slika 3. Vrednost ARI za nenormalizovane *Affymetrix* podatke, primenom algoritma k-srednjih vrednosti sa korelacijom

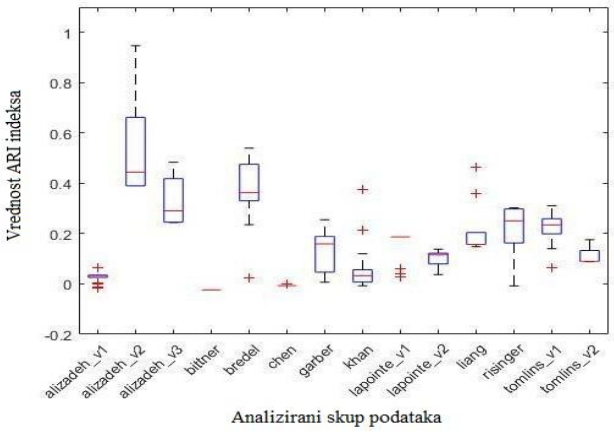


Slika 4. Vrednost Calinski – Harabasz indeksa za nenormalizovane *Affymetrix* podatke, primenom algoritma k-srednjih vrednosti sa euklidskim rastojanjem

Druga grupa podataka sadrži 14 skupova, koji su snimljeni pomoću cDNA mikročipa. Dobijeni rezultati ukazuju na značajnu zavisnost od početnih uslova i malu stabilnost rezultata (Slika. 5). Vrednosti ARI indeksa su uglavnom ispod 0,5. Dodatna normalizacija podataka ne dovodi do značajnog poboljšanja rezultata (Slika. 6).



Slika 5. Vrednost ARI za nenormalizovane *cDNA* podatke, primenom algoritma k-srednjih vrednosti sa euklidskim rastojanjem



Slika 6. Vrednost ARI za z-normalizovane *cDNA* podatke, primenom algoritma k-srednjih vrednosti sa euklidskim rastojanjem

IV. ZAKLJUČAK

Ovaj rad prikazuje rezultate klasterizacije dobijene primenom algoritma k-srednjih vrednosti nad obeležjima u svim skupovima podataka. Oni ukazuju na nemogućnost univerzalnog izbora parametara nad skupovima genskih ekspresija. Utvrđen je nedostatak robusnosti algoritma k-srednjih vrednosti nad većinom podataka i potreba za daljim unapređenjem postupka. Primećeno je da postoji potreba za efikasnijim tehnikama analize obeležja kao i analize koje bi ukazale na adekvatan izbor parametara za svaki skup podataka zasebno. Buduće analize će svakako uključiti niz različitih metrika, različite postupke normalizacije, ali i naprednije tehnike klasterizacije.

LITERATURA

[1] "The Top 10 Causes of Death", WHO, 2017.

[2] Z. Kakushadze, W. Yu, "K-means and cluster models for cancer signatures", *Biomolecular Detection and Quantification*, vol. 13, pp. 7-31, 2017.

[3] V. G. Tusher, R.Tibshirani, G. Chu, " Significance analysis of microarrays applied to the ionizing radiation response", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 5116-5121, 2001.

[4] I. Šašić, S. Brdar, T. Lončar – Turukalo, H. Aidos, A. Fred, "Consensus Clustering for Cancer Gene Expresision Data", BIOSTEC, vol. 3: bioinformatics, pp. 176 – 183, 2017.

[5] M. de Souto, I. Gosta, D. de Araujo, T. Ludemir, A. Schilep, "Clustering cancer gene expression data: a comparative study", BMC, pp. 9-497, 2008.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 14863-14868, 1998.

[7] S. Lloyd, "Least squares quantization in PCM". *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982.

[8] H. Steinhaus, "Sur la division des corp materiels en parties". *Bulletin of Acad. Polon. Sci.*, pp. 801–804, 1956.

[9] L. Hubert, P. Arabie, "Comparing partitions". *Journal of classification*, vol. 2, no. 1, pp. 193-218, 1985.

[10] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[11] T. Calinski, J. Harabasz, "A dendrite method for cluster analysis", *Communications in Statistics*, vol. 3, pp. 1-27, 1974.

ABSTRACT

Clustering algorithms are a form of unsupervised learning technics. They try to determine the structure of datasets and discover unknown bonds that exist between objects of interest. These methods are often used in order to group gene expressions, target therapies and discover new subtypes of cancer. Unfortunately, the large number of genes, as well as the presence of noise during the data acquisition make this process very difficult. This paper represents the evaluation of the k-means algorithm in clustering cancer gene expression data. The analysis is conducted using the original and the normalized form of the datasets, including two types of distances (Euclidean distance and correlation). The validation of the whole procedure is done using the internal (Silhouette and Calinski-Harabasz) and external measures of validation (Adjusted Rand Index). Based on these indices it can be concluded that on different datasets the results of the k-means algorithm show a large variability. The lack of robustness and the high sensitivity to parameter selection make this task worth exploring in the future.

Evaluation of the k-means Algorithm in Clustering Cancer Gene Expression Data

Katarina Ćoćić, Boris Knežević