**Opis izabrane metode selekcije značajki**

Metoda selekcije značajki za ulazni skup podataka od N emisija i p značajki vraća podskup značajki za koji clusteriranje je najuspješnije.

Za clusteriranje skupa podataka koristi se metoda OPTICS koja grupira emisije ovisno o njihovoj gustoći unutar p-dimenzionalnog prostora značajki. Gustoća emisija računa se pomoću parametara **epsilon** i **minNumPoints**. Epsilon određuje maksimalnu udaljenost između emisija u p-dimenzionalnom prostoru da bi pripadale istom clusteru. MinNumPoints određuje minimalni broj točaka u epsilon-susjedstvu od emisije kako bi točka pripadala određenom clusteru.

Uspješnost clusteriranja određuje se raznim mjerama grupacije emisija unutar clustera nazvanim indeksima clusteriranja. Najčešće indeksi su bazirani na omjeru udaljenosti točaka istog clustera od centroida (srednje vrijednosti) tog clustera te udaljenosti centroida različitih clustera (ponekad i globalnog centroida). Vrijednost indeksa određuje uspješnost clusteriranja gdje maksimizacija ili minimizacija vrijednosti kod određenih indeksa znači uspješnije clusteriranje.

Kako selekcijom značajki želimo dobiti skup značajki koji možemo vizualno prikazati podskupovi značajki ograničeni su na veličine 3 do 5 značajki. Isto tako od početnog skupa podataka od 21 značajki uzeti je podskup od 13 najbitnijih značajki prikazanih slikom ispod.

|  |  |  |  |
| --- | --- | --- | --- |
| **Značajka** | **Odabrani broj** | **Skaliranje vrijednosti s obzirom na min-max** | |
| **Min** | **Max** |
| **rise time** | 1 | 0 us = 0 | 400 us = 1 |
| **counts to** | 2 | # 0 = 0 | # 400 = 1 |
| **counts from** | 3 | # 0 = 0 | # 400 = 1 |
| duration |  | 0 us = 0 | 3000 us = 1 |
| **peak amplitude** | 4 | 0 V = 0 | 3 uV = 1 |
| average frequency |  | 1 kHz = 0 | 10 kHz = 1 |
| rms |  | 0 = 0 | 1 = 1 |
| asl |  |  |  |
| reverbation frequency |  |  |  |
| initial frequency |  |  |  |
| **signal strength** | 5 |  |  |
| absolute energy |  |  |  |
| **pp1** | 6 |  |  |
| **pp2** | 7 |  |  |
| **pp3** | 8 |  |  |
| **pp4** | 9 |  |  |
| **centroid frequency** | 10 |  |  |
| **peak frequency** | 11 |  |  |
| **max frequency** | 12 |  |  |
| num of freq peaks |  |  |  |
| **weighted peak frequency** | 13 |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Značajka** | **Skaliranje vrijednosti s obzirom na min-max** | | | |
| **Realni Min** | **Realni Max** | **Harcodirani Min** | **Harcodirani Max** |
| **RISE TIME** | 2.5 us | 377.5 us | 0 us = 0 | 400 us = 1 |
| **COUNTS TO** | 4 | 279 | # 0 = 0 | # 400 = 1 |
| **COUNTS FROM** | 2 | 194 | # 0 = 0 | # 400 = 1 |
| **DURATION** | 25.5 us | 424 us | 15 us = 0 | 500 us = 1 |
| **PEAK AMPLITUDE** | 0.19815 V 0.28739 mV | 198.15 mV | 0 V = 0 | 200 mV = 1 |
| **AVERAGE FREQUENCY** | 0.7 MHz | 1.77 MHz | 0.1 MHz = 0 | 2 MHz = 1 |
| **RMS** | 35.25 uV | 22.376 mV | 0 V = 0 | 100 mV = 1 |
| **ASL** | -102.845 dB | 44.1081 dB | -100 dB = 0 | 100 dB = 1 |
| **REVERBATION FREQUENCY** | 0.346 MHz | 2 MHz | 0.1 MHz = 0 | 2 MHz = 1 |
| **INITIAL FREQUENCY** | 0.16 MHz | 2,4 MHz | 0.1 MHz = 0 | 2 MHz = 1 |
| **SIGNAL STRENGTH** | 3.6899 nVs | 3.1905 uVs | 0 nVs = 0 | 10 uVs = 1 |
| **ABSOLUTE ENERGY** | 6.362e-5 p aJ | 25.635 p aJ | 0 aJ = 0 | 100 p aJ = 1 |
| **PP1** | min= 2.8035 u% | 0.064765 % | 0 % = 0 | 100 % = 1 |
| **PP2** | 0.023486 % | 0.96706 % | 0 % = 0 | 100 % = 1 |
| **PP3** | 0.02384 % | 0.94553 % | 0 % = 0 | 100 % = 1 |
| **PP4** | 8.4381e-05 % | 0.63402 % | 0 % = 0 | 100 % = 1 |
| **CENTROID FREQUENCY** | 0.199 MHz | 0.473 MHz | 0.1 MHz = 0 | 1 MHz = 1 |
| **PEAK FREQUENCY** | 0.117 MHz | 0.798 MHz | 0.1 MHz = 0 | 1 MHz = 1 |
| **MAX MAGNITUDE OF FREQUENCY SPECTRUM** | 6.9982 uV | 8.6415 mV | 0 V = 0 | 200 mV = 1 |
| **NUM OF FREQ PEAKS** | 0 | 7 | # 0 = 0 | # 10 = 1 |
| **WEIGHTED PEAK FREQUENCY** | 0.185 MHz | 0.614 MHz | 0.1 MHz = 0 | 1 MHz = 1 |

**DBscan algoritam: (Density-Based Spatial Clustering of Applications with Noise)**

[How Does DBSCAN Clustering Work? | DBSCAN Clustering for ML (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/)

Služi za grupiranje skupa podataka na temelju njihove gustoće, a ne oblika. Isto tako otporan je na outliere te nije potrebno zadati broj clustera kao parametar. Za računanje potrebna su mi dva parametra:

1. Epsilon – određuje radijus kružnice koji se kreira oko svake točke skupa podataka kako bi se odredila gustoća. U višim dimenzijama kružnica postaje sfera i tako dalje.
2. minPoints – minimalni broj točaka unutar kružnice koju definira epsilon da bih ta točka bila labelirana kao Core točka

Klase točaka određene na temelju navedenih parametara su slijedeće:

1. Core točka – točka koja na udaljenosti Epsilon od sebe ima barem broj točaka minPoints -1
2. Border točka - točka koja na udaljenosti Epsilon od sebe ima manje točaka od minPoints-1, ali ima barem jednu točku
3. Noise točka - točka koja na udaljenosti Epsilon od sebe ima 0 točaka

Klase parova točaka x i y određene su kao:

1. Direktno Density-Reachable (DDR) – kada točka x se nalazi u susjedstvu od y te y je Core točka. Zapis: (x,y) = DDR
2. Density-Reachable (DR)– kada postoji lanac direktno density-reachable točaka koji kreću od y prema x. Npr. (x,y) = DR ako (p1,y) = DDR i (p2,p1) = DDR i (x,p2) == DDR
3. Density-Connected (DC) – Ako postoji točka o koja je DR s točkama x i y.

Npr. (x,y) = DC Ako (x,o) = DR i (y,o) = DR

Za računanje udaljenosti između točaka koristi Euklidsku udaljenost iako je moguće i druge metrike.

Algoritam treba proći kroz čitavi skup podataka **samo jedanput**. (O(N))

Izbor parametara:

1. **Epsilon >= Broj\_Dimenzija** + 1 ili Broj\_Značajki + 1
   1. Može se odrediti pomoću K-distance grafa gdje je optimalni Epsilon u točci s najvećom zakrivljenosti grafa (koljeno ili lakat grafa)
   2. Ako je Epsilon preveliki onda će biti napravljeno više clustera i više Noise točaka
   3. Ako je Epsilon premali onda će manji clusteri biti skupljeni u jedan veliki i izgubiti će se detaljnost
2. minPoints >= 3, Najčešće **minPoints = 2\*Broj\_Dimenzija**

**OPTICS algoritam:**

Za clusteriranje skupa podataka koristi se metoda OPTICS koja grupira emisije ovisno o njihovoj gustoći unutar p-dimenzionalnog prostora značajki. Gustoća emisija računa se pomoću parametara epsiloni minNumPoints. **Epsilon** određuje maksimalnu udaljenost između emisija u p-dimenzionalnom prostoru da bi pripadale istom clusteru tj. minimalnu udaljenost između različitih clustera. **MinNumPoints** određuje minimalni broj točaka u epsilon-susjedstvu od emisije kako bi emisija pripadala određenom clusteru tj. minimalni broj emisija potrebnih da čine cluster.

Reachability distance = RD

Pomoću RD možemo zaključiti

Indeksi koji si koriste i značenje:

Davies-Bouldin (DB)

**Silhouette**

Razina vrijednosti od -1 do 1. Negativne vrijednosti označuju da su točke krivo raspoređene, dok gdje pozitivne da su točke dobro raspoređene (0 je neutralno). Svakoj točci dodjeljuje se silhouette vrijednost koja mjeri koliko je slična ta točka točkama u istom clusteru, uspoređujući s točkama u drugim clusterima. Ako je točka samostalan cluster onda je njezina silhouette vrijednost jednaka 1. (U literaturi piše na 0)

a(i)= srednja vrijednost udaljenosti između točke i te svake ostale točke unutar istog clustera

b(i,k)= srednja vrijednost udaljenosti između točke i te svake točke unutar različitog clustera k

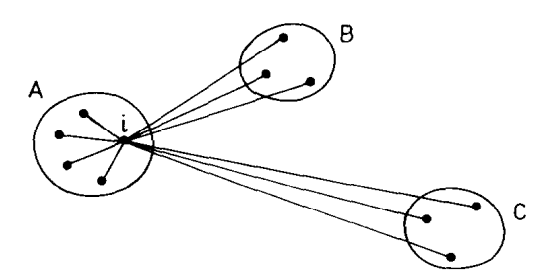
= d(i, C)

b(i) = minimalna b(i,k) tj. srednja vrijednost udaljenost točaka clustera najbliže točci i i točke i

= min d(i,C)

Cluser B koji za koji je vrijednost d(i,C) najmanja nazivamo susjed točke i. Ovo je kao drugi najbolji izbor clustera za točku i, gdje je trenutni klaster točke i prvi izbor. Silhouette vrijednost točke i određuje koliko dobro točka i pripada svojem clusteru.

s(i) =



Ukupni silhouette indeks određuje srednju vrijednost pripadnosti svake točke svojem clusteru.

S = srednja vrijednost svih s(i)

**Calinski–Harabasz**

Još se naziva kriterij omjera varijance (VRC). Varijanca je ovdje definirana kao suma kvadrata međusobnih udaljenosti skupa točaka pomnoženih sa brojem točaka. Nisam siguran zašto se množi, a ne dijeli s brojem točaka.

SSb = varijanca udaljenosti centroida clustera od centroida čitavog skupa podataka.

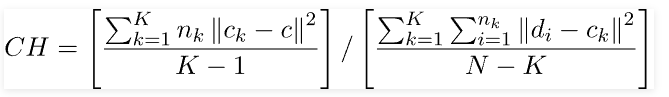
= dobro definirani klasteri imaju visoku varijanca udaljenosti između klastera

SSw = sveukupna varijanca unutar klastera, varijanca udaljenosti centroida clustera od točaka tog klastera

= dobro definirani klasteri imaju nisku varijanca udaljenosti unutar klastera

Bitan je broj klastera.

CH = , N – broj točaka, k – broj klastera



CH kriterij je najprikladniji u slučaju kada su klasteri manje-više sferni i kompaktni u svojoj sredini (kao što su npr. normalno raspoređeni) Pod jednakim ostalim uvjetima, CH preferira klaster rješenja s klasterima koji se sastoje od otprilike istog broja objekata. [[r - What is an acceptable value of the Calinski & Harabasz (CH) criterion? - Cross Validated (stackexchange.com)](https://stats.stackexchange.com/questions/52838/what-is-an-acceptable-value-of-the-calinski-harabasz-ch-criterion)]

**generalized Dunn**

[[Dunn index - Wikipedia](https://en.wikipedia.org/wiki/Dunn_index)]Dunn indeks omjer minimalne inter-cluster udaljenosti i maksimalne intra-udaljenosti.

Intra-cluster udaljenost za određeni cluster se može računati kao:

1. Maksimalna udaljenosti između svih parova točaka unutar istog clustera
2. Srednja vrijednost udaljenosti između svih parova točaka unutar istog clustera
3. Srednja vrijednost udaljenosti između točaka i centroida clustera unutar istog clustera

Maksimalna intra-cluster udaljenost se uzima kao najgori raspoređeni klaster.

Inter-cluster udaljenost između dva clustera se može računati kao:

1. Udaljenost između najbližih točaka clustera
2. Udaljenost između najdaljih točaka clustera
3. Srednja vrijednost udaljenosti između svih točaka dvaju clustera
4. Udaljenost između centroida clustera
5. Srednja vrijednost udaljenosti svih točaka i njihovog centroida clustera

Minimalna inter-cluster udaljenost se uzima kao najgori odvojeni klasteri.

Maksimizacijom Dunn indeksa znači da su točke unutar klastera bliže raspoređene, a klasteri su međusobno više udaljeni.

Cilj je identificirati skupove klastera koji su kompaktni, s malom varijacijom između točaka clustera i dobro odvojeni, gdje su srednje vrijednosti različitih clustera dovoljno udaljene, u usporedbi s varijancom unutar klastera.

**COP - Context-independent Optimality**

Koristi se kod hijerarhijskog clusteriranja.

Intra-variance – srednja vrijednost udaljenosti između točaka i centroida clustera unutar istog clustera

Inter-variance - Udaljenost između najdaljih točaka clustera

**COP** = srednja vrijednost omjera intra-variance i minmalnog inter-variance za svaki cluster pomnožena sa brojem točaka unutar klaster (varijanca)

Gamma

Tou

**Rezultat:**

**Tou** result are features: **rise time counts to pp3**

**Silhouette** result are features: **rise time peak amplitude max frequency**

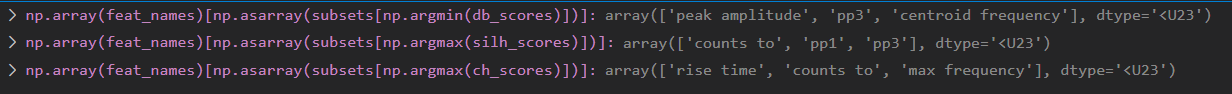
**Gamma** result are features: **peak frequency max frequency weighted peak frequency**

**DB** result are features: **signal strength centroid frequency max frequency**

**Calinski** result are features: **peak frequency max frequency weighted peak frequency**

[1] - [2016, Vergeynst, Clustering reveals cavitation-related acoustic emission signals from dehydrating branches.pdf](file:///C:\Users\bujak\Desktop\FER\5.%20godina\DIPLOMSKI%20PROJEKT\DILPOMSKI%20RAD\Review%20poslane%20literature\2016,%20Vergeynst,%20Clustering%20reveals%20cavitation-related%20acoustic%20emission%20signals%20from%20dehydrating%20branches.pdf)

Koristeći OPTICS iz python-a za iste značajke i indexe DB, CH te Silh dobio sam skupove značajki:



Max silh score: 0.06054845328549051

Max db score: 1.2018091586107562

Max ch score: 22.431649222979225

silh feats: ['pp1' 'peak frequency' 'max frequency']

db feats: ['counts to' 'peak amplitude' 'pp2' 'pp3' 'pp4']

ch feats: ['rise time' 'counts to' 'max frequency']

Voting scheme 1st features: (0, 1, 4, 13, 14)

Voting scheme 2nd features: (0, 1, 4, 14, 15)

Voting scheme 3rd features: (0, 1, 13, 15)

Voting scheme 4th features: (0, 1, 14, 15, 18)

Voting scheme 5th features: (0, 1, 12, 14)

Implementacija OPTICS algoritam funkcionira dobro u matlabu za dobro izabrane parametre **Epsilon** i **MinNumPoints.** Međutim potrebno je za svaki podskup značajki kvalitetno odrediti navedene parametre kako bi uspijeli usporediti rezultate clusteriranja svakog podskupa značajki.

Najveći mogući epsilon bih bila ujedno i najveća udaljenost između točaka u p-dimenzionalnom prostoru što se računa po formuli (1).

max\_euclid\_dist = euclid\_dist(min\_val,max\_val) = (p\*(min-maxval))^(1/2)

Kako je skup podataka normaliziran na vrijednosti od 0 do 1 (opisano u prijašnjem poglavlju) tada je najveća udaljenost je p^(1/2) tj. 2.2361. Za svaku 3 dimenzije jednaka je 1.7321.

Kako **MinNumPoints** se određuje …

Testiranjem raznih epsilon parametara i min\_points nad podskupovima značajki od 3 značajke vizualno ispada mi najbolje clusteriranje za Epsilon vrijednost od 0.2 do 0.4 te MinNumPoints od 10 do 20.  Za oba skupa podataka Epsilon je najbolji u navedenom interval, dok za manji skup podataka clustering daje najbolje rezultate kada je MinNumPoints od 10 do 12, a za veći skup podataka 15 do 20. Navedeno je vidljivo iz figura na lokaciji figures/clustering/test\_optics\_params/.

Uzimajući u obzir navedene parametre napravio sam feature selection za sve podskupove od 3 značajke s početnim skupom značajki:

(rise time, counts to, counts from, peak amplitude, signal strength,pp1, pp2, pp3, pp4, centroid frequency, peak frequency, max frequency, weighted

peak frequency)

za sve kombinacije ulaznih parametara za Epsilon 0.2, 0.3 ili 0.4 te MinNumPoints 10, 11 ili 12.

Za cluster validity index koristio sam najjednostavniji gdje se intra-distance clustera računa kao srednja udaljenost od centroida clustera, inter-distance clustera se računa kao udaljenost između centroida clustera. Omjerom te dve udaljenosti dobio sam iznos indeksa. Indeks sam još pomnožio sa postotkom outliers-a unutar skupa podataka kako bez toga dijela feature selection je preferirao male clustere s puno outliersa.

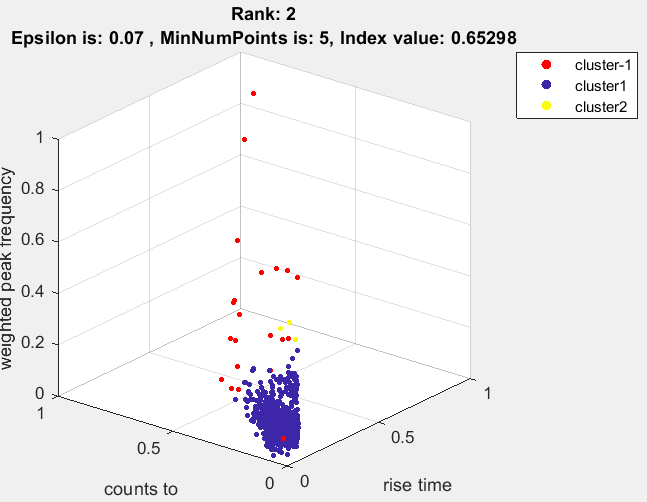
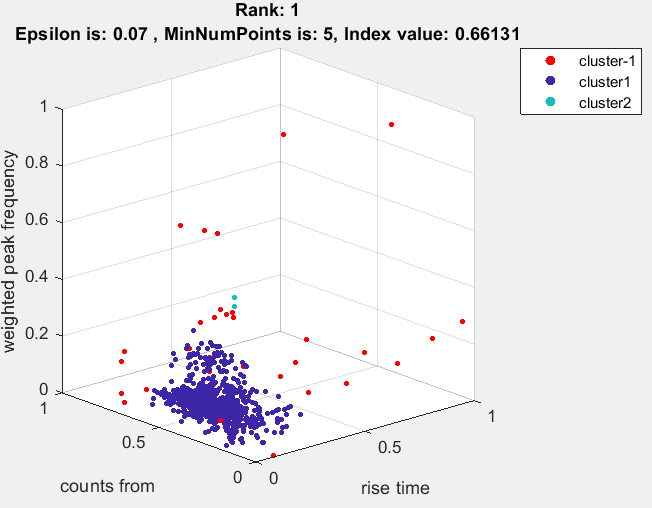
Što je vrijednost indeksa veća to je clusteriranje bolje.

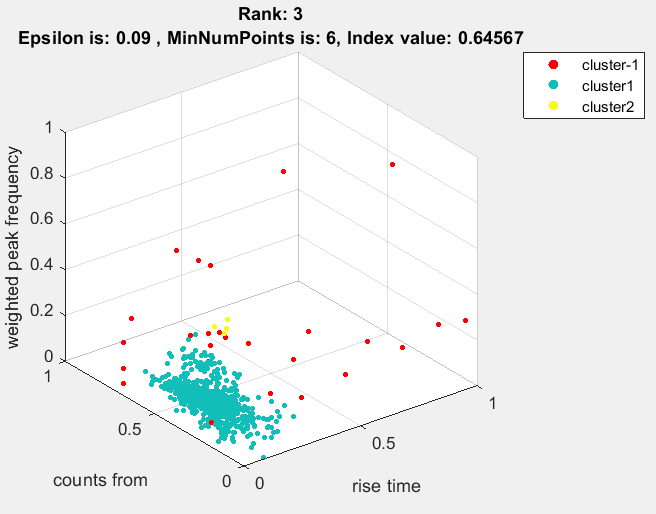
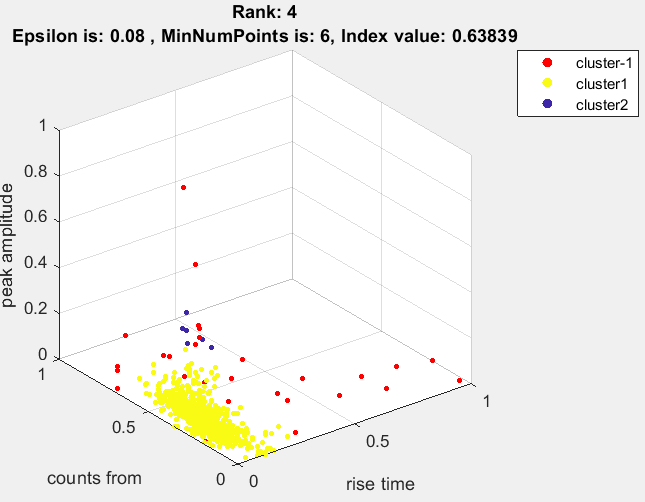
Prvih 5 najboljih rezultata su prikazana u tablici i slikama ispod. Outliers-i su prikazani crvenom bojom.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Poredak** | **Značajke** | | | **Epsilon** | **MinNumPoints** | **Vrijednost indeks** |
| 1. | Weighted peak frequency | Peak frequency | Rise time | 0.02 | 10 | 15.6898 |
| 2. | Peak frequency | Signal strength | PP1 | 0.03 | 12 | 14.004 |
| 3. | Peak frequency | Signal strength | Rise time | 0.04 | 12 | 12.8092 |
| 4. | Peak frequency | Counts to | PP1 | 0.02 | 10 | 12.3785 |
| 5. | PP4 | Counts to | Signal strength | 0.04 | 10 | 11.4746 |

**19.5.2022.**

Proba s selekcije značajki s MinNPoints od 5 do 20 daje ogromni jedan cluster, dok drugi je samo 5-7 točaka. Zaključak da je potrebno povećati broj MinNPoints na 20 do 40 pa analizirati rezulatate.



20.5.2022.

Keil compiler error:

[ARM: Warning L6989W](https://developer.arm.com/documentation/ka002865/latest)

[LICENSE: L6050U: The Code Size of this Image Exceeds the Maximum ... (arm.com)](https://developer.arm.com/documentation/ka002709/latest)