**Review poslane literature**

# 2012, Sause, Pattern recognition approach to identify natural clusters of acoustic emission signal

Kaže kako diskriminaciju akustične emisije i šum je moguće postići s nenadziranim tehnikama za prepoznavanje uzoraka (*unsupervised pattern recognition*).

Za procjenu kvalitete grupiranja dataseta navedena je nekolicina indeksa valjanosti:

(izabrani zbog male kompleksnosti)

**Davies–Bouldin indeks – *DB* – bolje grupiranje kod minimizacije indeksa**

**Definicija:** [Davies–Bouldin index - Wikipedia](https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index)

Prvo se određuje unutarnje raspršivanje svakog klastera ***i*** kao tj. srednja udaljenosti točaka unutar klastera od centroida klastera. Za mjeru udaljenosti najčešće se koristi Euklidska udaljenost, iako je važno da mjera udaljenosti se računa identično kao mjera udaljenosti za grupiranje klaster kako bi rezultanti indeks bio značajan. Što je vrijednost manja to su klasteri gušće grupirani tj. bolje grupirani.

Nakon toga se za svaki par klastera ***i*** i ***j*** određuje kao mjera separacije između klastera ***i*** i ***j***. Za mjeru separacija klastera najčešće se koristi Euklidska udaljenost između centroida odgovarajućih klastera. Što je vrijednost veća to su klasteru udaljeniji tj. bolje grupirani.

Zatim za svaki par klastera ***i*** i ***j*** određuje kao mjera sličnosti klastera. Računa se kao omjer udaljenosti između centroida 2 klastera (**)** te zbroja unutarnjeg raspršivanja svakog od 2 klastera (). Što je vrijednost veća to su 2 klastera sličnija (bliži međusobno i interno raspršeniji) tj. gore grupirani.

Za završnu vrijednost **Davies–Bouldin indeksa** računa se kao srednja vrijednost najveće mjere sličnosti svakog klastera tj. mjera sličnosti klastera gdje jer par klastera najsličniji. **Što je vrijednost Davies–Bouldin indeksa manja to je grupiranje bolje izvedeno.**

**Matlab implementacija:** [Evaluate clustering solutions - MATLAB evalclusters - MathWorks Nordic](https://se.mathworks.com/help/stats/evalclusters.html)

Matlab implementacija je izvedena unutar funkcije *evalclusters* gdje je potrebno zadati opciju 'DaviesBouldin' kao mjeru izračuna grupiranja.

**Tou indeksi – 𝜏 – bolje grupiranje kod maksimizacije indeksa**

**Rousseeuwova metoda provjere siluete - *S*– bolje grupiranje kod maksimizacije indeksa**

**Hubertova gama statistika – Γ – bolje grupiranje kod maksimizacije indeksa**

Budući da svaki indeks valjanosti klastera pokazuje drugačije performanse ovisno o pojavi izvanrednih vrijednosti (***outliers***) i oblik klastera, kombinirana procjena višestrukih indeksa validacije klastera stvara veću mogućnost povećanja pouzdanosti metoda identifikacije klastera. Takvu kombiniranu evaluaciju već su predložili Günter i Bunke (2003.) [1] u obliku glasovanja te **se koristi unutar ovog rada za selekciju značajki**.

**Metodologija prepoznavanja uzoraka**

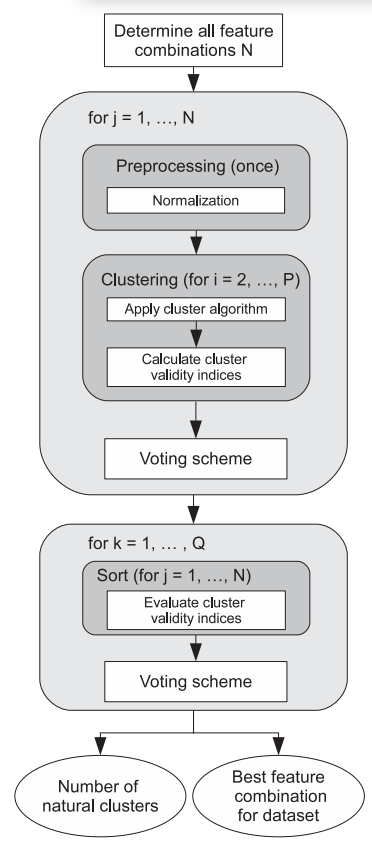
Cilj metodologije je uvesti tehniku koja automatski izabire obećavajuće značajke za grupiranje akustičnih emisija. Ovakva iscrpna metoda pretraživanja skupa značajki ovisi o kvaliteti prvobitnih značajki te vremenski o veličini skupa značajki.

Prvi korak je izbor svih podskupova značajki koji sadrže minimalno M značajki. Broj ukupnih kombinacija tj. podskupova nad kojima će se provesti grupiranje je N te broj značajki u svakom podskupu je K.

Svi slijedeći koraci rade se na svakom od N podskupova značajki posebno:

1. **Normalizacija** značajki radi se dijeljenjem sa standardnom devijacijom značajke što je bazirano na prijašnjim istraživanjima klasificiranja akustičnih emisija u [2], [3] i [4] gdje navode [5].
2. Višestruko **Grupiranje** k-means algoritmom tako da se svaki put zada drukčiji broj klustera od 2 do ***P***.
3. **Izračun indeksa valjanosti grupiranja** spomenutih posebno za rezultate svakog grupiranja s drukčijim brojem zadanih klastera.
4. Kod **sheme glasanja** **za izbor broj klastera,** klaster koji je imao najbolji indeks valjanosti grupiranja dobiva ***P*** bodova, broj klastera s drugim najboljim indeksom valjanosti grupiranja dobiva P-1 bodova te tako sekvencijalno do broja klastera s najgorim indeksom valjanosti koji dobiva 2 boda. Bodovi se dijele za svaku metodu računanja indeksa valjanosti. Broj klastera koji konačno ima najviše bodova se izabire kao optimalan.

Kod **sheme glasanja** **za izbor najboljeg podskupa značajki** se podskupu značajki s najboljim indeksom valjanosti grupiranja dodjeljuje 25 bodova, drugom najboljem 24 boda te tako sekvencijalno do najgoreg podskupa značajki. Izabrano je 25 bodova kako ima 4 mjere indeksa valjanosti te time podskup značajki može optimalno dobiti 100 bodova.



Slika Metodologija izbora optimalnog skupa značajki i broja klastera za grupiranje akustičnih emisija

**Umjetno generirani skup podataka**

Metodologija izbora značajki, broja klastera te kvaliteta grupiranja provjerena je na 54 umjetno generirana skupa podataka za koje je već poznato broj klastera, optimalne tj. ne šumovite značajke te kojem klasteru pripadaju te Rand indeks. Metodologija je svakom slučaju izabrala ispravan broj klastera te optimalan broj značajki. Isto tako Rand indeks je varirao od 0.748 do 1.

**Skup podataka akustičnih emisija**

Koristilo se 5 skupova podataka akustičnih emisija gdje je se računalo za sve podskupove značajki od 12 značajki do 5. Što daje N = 3302 kombinacija značajki za svaki skupo podataka. Korišteno je 12 značajki opisanih tablicom ispod:

Slika na kojoj se prikazuje tekst, stol

Opis je automatski generiran

**Ja u svojem radu ne koristim weighted peak-frequency i drukčije su mi raspoređeni partial powers.**

Rezultati na 6 dataseta na optimalan **broj klastera od 3 do 4** te broj **optimalnih značajki je 5**, a najvažnije su iz frekvencijske domene. Izbor značajki i broja klastera te indekse valjanosti grupiranja mogu se vidjeti na slici ispod:

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Podskupovi značajki koji su mogli optimalno razdvojiti klastere u vizualno odvijene grupe prikazane u dimenzijama određenih značajki su označeni kao Suitable. Glavne značajke koje prikazuju optimalno odvajanje skupa podataka na grupe su **weighted peak-frequency** i **partial power 2**.

Kako metodologija nije izdvojila uvijek isti podskup značajki kao optimalni, potrebno je bilo izračunati optimalni podskup značajki za sve skupove podataka. To se učinilo spajanjem svih skupova podataka u jedan te se metodologija provela na njemu u 10 iteracija. Rezultat je prikazan tablicom ispod:

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Zaključuje se kako su najbolji podskupovi značajki s ID-om 85 i 110 te se izabere 110 jer kada je metodologija izabrala podskup značajki s ID-om 85 onda je drugi najbolji bio ID 110.

Što znači optimalne značajke za grupiranje akustičnih emisija su:

* **Partial power 1**
* **Partial power 2**
* **Partial power 4**
* **Peak Frequency**
* **Weighted peak frequency**

Za kraju kaže članak da se metodologija može poboljšati ako se koristi hijerarhijski algoritam grupiranja.

[1] - [Sci-Hub | Validation indices for graph clustering. Pattern Recognition Letters, 24(8), 1107–1113 | 10.1016/s0167-8655(02)00257-x](https://sci-hub.se/https:/doi.org/10.1016/S0167-8655(02)00257-X)

[2] – [(PDF) Acoustic emission investigation of coating fracture and delamination in hybrid carbon fiber reinforced plastic structures (researchgate.net)](https://www.researchgate.net/publication/261993043_Acoustic_emission_investigation_of_coating_fracture_and_delamination_in_hybrid_carbon_fiber_reinforced_plastic_structures) , Lokalno: [Sause2008-JAE-Vol26.pdf](file:///C:\Users\bujak\Desktop\FER\5.%20godina\DIPLOMSKI%20PROJEKT\DILPOMSKI%20RAD\Review%20poslane%20literature\Dodatno%20iz%20poslanih%20tekstova\Sause2008-JAE-Vol26.pdf)

[3] - [Sci-Hub | Quantification of metallic coating failure on carbon fiber reinforced plastics using acoustic emission. Surface and Coatings Technology, 204(3), 300–308 | 10.1016/j.surfcoat.2009.07.027](https://sci-hub.se/https:/doi.org/10.1016/j.surfcoat.2009.07.027)

[4] - [Influence of specimen geometry on acoustic emission signals in fiber reinforced composites: FEM-simulations and experiments (ndt.net)](https://www.ndt.net/article/ewgae2010/papers/49_Sause.pdf)

[5] - Anastassopoulos A., Philippidis T., Journal of Acoustic Emission 13, 1995, p.11

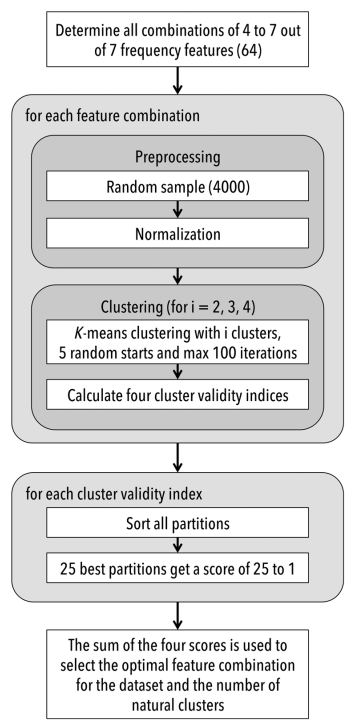
# 2016, Vergeynst, Clustering reveals cavitation-related acoustic emission signals from dehydrating branches

Unutar članka obrađuje se pitanje klasifikacije AE emisija nastale tijekom eksperimenta dehidracije biljaka. Želi se odrediti koje AE emisije se poklapaju sa trenutcima kavitacije ili embolizacije koji označavaju nedostatak vode kod biljaka. Eksperimenti su detaljno provedeni i opisani od procesa nabava biljke do akvizicije AE emisija (frekvencija uzorkovanja = 10 MHz) te raznih mjerenja kao što su VWC [kg m-3] (volumni vodeni sadržaj), ψ [MPa] (vodeni potencijal ksilema) i VC (Krivulje ranjivosti) (Spominje se i eksperiment koji koristi mikro računalnu tomografiju (µCT) ). Sva mjerenja omogućuju razumno labeliranja AE emisija na temelju fizioloških značajka biljke u danom trenutku.

Kod predprocesiranja izračunato je 11 značajki:

* Amplitude (dB)
* Rise Angle (rad)
* Rise time (µs)
* Apsolutna energija (aJ)
* Peak frequency (PF) (Hz)
* Frequency centroid (FC) (Hz)
* Weighted PF (WPF) (Hz) – geometrijska srednja vrijednost od PF i FC
* Partial power 1 (PP1) (%) – 0-100 kHz
* Partial power 2 (PP2) (%) – 100-200 kHz
* Partial power 3 (PP3) (%) – 200-400 kHz
* Partial power 4 (PP4) (%) – 400-800 kHz

Koriste grupiranje bazirano na prijašnje spomenutoj metodologiju uz manje izmijene prikazane slikom ispod. Broj klastera je bio ograničen na 2 do 4 te broj značajki u optimalnom podskupu značajki je bio ograničen na 4 do 7. Ukupno je bilo N = 64 mogućih podskupova značajki. Uzeti su 4000 slučajnih uzoraka iz kombiniranih skupova podataka jer bi inače vrijeme izračuna bilo preveliko.



Pronađeno je distinktivna 2 klastera s podskupom značajki FC, PF, WPF i PP4. Prvi klastera od 2% emisija ima emisije s visokim WPF i PP4, dok drugi klaster emisije imaju velike varijacije u pojasu 100-200 kHz. Nakon toga napravljeno je grupiranje drugog većeg klastera koji se podijelio u 2 pomoću značajki WPF, PF, PP1 i PP2. Bazirano na frekvencijskom sadržaju 3 klastera su labelirana nisko, srednje i visoko frekvencijski klasteri. Nisko i srednje frekvencijski klasteri su imali visoki PP1 i PP2, dok visoko frekvencijski klaster je imao visoke PP4, WPF, PF i FC. Srednje frekvencijski klasteri su grupirani u PF značajci oko 200 kHz.

**Zaključci o klasterima**

Nadalje istražili su jesu li pronađeni klasteri pripadaju određenom fiziološkom procesu tijekom dehidracije pomoću mjerenja spomenutih u početku.

Srednje frekvencijski klaster bio je najviše povezan s kavitacijom. Visoko frekvencijski klaster vjerojatno je rezultat kapilarnog djelovanje slobodne vode i brze kontrakcije kore tijekom elastičnog skupljanja, dok je nisko frekvencijski klaster bio najvjerojatnije uzrokovano makro-frakturama ili Hainesovim skokovima.

# 2021, Muir, Nature, Damage mechanism identification in composites via machine learning and acoustic emission

Kaže kako se ne zna je li k-means algoritam prikladan primjeni na AE emisijama jer bi se trebao primjenjivati na izotropnim (isti vrijednosti u svim smjerovima) skupovima podataka, dok za AE emisije ne znamo kakva svojstva imaju grupe.

Članak se bavi klasificiranjem načina oštećenja pomoću algoritma strojnog učenja na skupo podataka AE emisija nastalih prilikom oštećenja. Za generiranje AE signala moguće je koristiti određene modele koje simuliraju eksperimentalni nastanak signala dosta dobro.

Opisuje se nenadzirano učenje kao: (**podebljano se obrađuje u članku**)

1. Eksperiment gdje se prikupe n uzoraka
2. **Ekstrakcija d značajki iz skupa uzoraka**
3. **Izabrana je metrika značajki te si radi potrebna transformacija značajki**
4. **Izabran algoritam strojnog učenja (grupiranje)**
5. Analiza rezultata grupiranja

**Clustering = pattern recognition ?**

Opisuje se korištenje i izračun waveleta koji opisuje vremensku i frekvencijsku domeni pod nazivom kontinuirani wavelet transform (CWT). Isto tako opisuju se ostale vrste waveleta kao WPT i HHT koje bi mogle zamijeniti FFT kao značajka.

Opisuje kako PCA može biti korisna međutim da je nepotrebna za AE klasifikaciju kako je najčešće broj značajka manji. Međutim PCA izbjeljenje (whitening ili sphering) može pomoći popraviti značajke. PCA whitening ponovno skalira PCA osi tako dijagonalna kovarijacijska matrica dobivena PCA postane matrica identiteta. Opisuje k-means, **fuzzy c-means i Gaussian mixture model** algoritme.

Opisuje problem određivanja broja klastera pomoću metoda vrijednost Siluete, Akaike information criterion (AIC) i Bayesian information criterion (BIC).

Opisuje samo-oragnizirajuće mape (SOM) i t-SNE. Služe za vizualnu reprezentaciju visoko dimenzionalnih podataka tj. zapravo redukcija značajki.

Opisuje potrencijalne probleme.

# A\_review\_of\_feature\_selection\_methods\_with\_applications

Koraci selekcije značajki:

1. Generiranje podskupa značajki
2. Evaluacija podskupa značajki – evaluacijski kriterij
3. Ponavljanje 1. – 2. dok se ne postigne kriterij obustave
4. Kriterij obustave evaluacije – izabran najbolji podskup značajki
5. Validacija rezultata

Podjela selekcije značajki po metodama izvođenja:

* **Filter metoda** – određuju kvalitetu značajki neovisno o algoritmu klasifikacije skupa podataka

**Univarijantne** metode - Rangiraju individualne značajke

**Multivarijantne** metode - Evaluiraju podskup značajki

Za izbor podskup značajki potrebna je **strategija pretraživanja**:

1. Naprijedna – dodavanje značajki i prazan skup
2. Nazadna – uklanjanje značajki iz punog skupa
3. Dvosmjerna – pretražuje i prazan skup te i puni skup značajki
4. Heuristička – Generira početni podskup značajki pomoću heurističkog (genetičkog) algoritma te obrađuje ga

Brzina promjene skupa značajki strategije pretraživanja:

1. Eksponencijalno – podskup značajki raste eksponencijalno
2. Sekvencijalno – može dovesti do lokalnog minimuma
3. Slučajno – uklanja mogućnosti lokalnog minimuma

Podjela metoda po **mjeri filtriranja** značajki na:

1. Informacija
2. Udaljenost
3. Konzistencija
4. Sličnost
5. Statistička mjerenja

Podjela metoda ovisno o **tipu skupa podataka**:

1. Klasifikacija
2. Regresija
3. Grupiranje (Klastering)
4. ~~Semi – nadgledano učenje~~ (Ne obrađuje se) Daje dodatnu literaturu: [Sci-Hub | Discriminative Semi-Supervised Feature Selection Via Manifold Regularization. IEEE Transactions on Neural Networks, 21(7), 1033–1047 | 10.1109/tnn.2010.2047114](https://sci-hub.se/https:/dl.acm.org/doi/abs/10.1109/TNN.2010.2047114.)

Prikaz potencijalnih metoda na slici ispod po spomenutim podjelama.

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Analiza metoda primjenjive kod grupiranja (clustering-a): [download (psu.edu)](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.295.8115&rep=rep1&type=pdf)

**Spektralna selekcija značajki (SPEC)**

SPEC proučava kako odabrati značajke prema strukturama grafa dobivenim iz matrice sličnosti. Glavna ideja iza SPEC-a je da su značajke u skladu sa strukturom grafa dodijelio slične vrijednosti instancama koje su blizu jedna drugoj na grafu. Zaključno te značajke trebaju biti relevantne jer se ponašaju slično u svakom klasteru. Motivirano teorijom grafova koja kaže da informacije o strukturi grafa mogu biti uhvaćene iz svog spektra.

SPEC je multivarijantna metoda selekcija značajki koja procjenjuje relevantnost značajke procjenom konzistentnosti značajke sa spektrom matrice izvedena iz matrice sličnosti ***S***. SPEC koristi funkciju radijalnih baza RBF kao funkciju sličnosti između dva uzorka i .

RBF je bilo koja funkcija čija vrijednost je udaljenost (radijalna) između ulazne točke i određene fiksne točke, to je najčešće Euklidska udaljenost. [1] Matrica sličnosti je matrica kod koje veće vrijednosti označavaju veću sličnost što je suprotno od matrice udaljenosti. (npr. korelacijska matrica) [2]

Iz matrice sličnosti dobivamo matricu susjedstva (*adjacency*) ***W*** te iz nje graf ***G*** kao KNN graf ili neki drugi. ***W*** se dobiva tako da za svaki red ili stupac (čvor) u matrici sličnosti uzmemo vrijednosti K najsličnijih parova tj. gdje je vrijednost najveća. Ostale vrijednosti postavimo u nulu. KNN graf se crta kao spajanje čvorova s težinom poveznice u matrici susjedstva s time da ne spajamo parove čvorova s vrijednosti nula kao što vidimo na slici ispod. [3]

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Iz ***W*** onda dobivamo kao matricu stupnjeva koja ima dijagonalu jednaku sumi redova ***W***.

Iz i ***W*** računa se Laplaceova matrica L i normalizirana Laplaceova matrica.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Težina svake značajke se računa pomoću određenih funkcija koje koriste Laplaceove matrice.

**Laplaceovi rezultat (Laplacian** **score)**

Slično kao **SPEC** ako se koristi funkcija rangiranja značajki: (fi – vektor značajke, ostali vektori opisani u SPEC)

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Algoritam SPEC koji koristi LS možemo vidjeti na slici ispod.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

**Odabir značajke za rijetko grupiranje (Feature selection for sparse clustering)**

Sparse clustering je klasteriranje kada broj podataka u skupu podataka je manji od broja značajki Prednosti metode je ako manje grupe se razlikuju samo u značajkama onda rezultati grupiranja mogu biti bolji.

Skup podataka se matričnom dekompozicijom dijeli na manji i veći dio. Tako da se aproksimira n×p matrica podataka X kao X ≈ AB gdje je A n×q matrica i B je q×p matrica, za q << p . Zatim se može grupirati promatranja koja koriste A kao matricu podataka, a ne X. Na primjer napraviti PCA od A i grupirati prvih nekoliko PC. Međutim učinak izvođenja PCA za smanjenje podataka dimenziju prije grupiranja nije opravdano budući da glavne komponente s najvećim svojstvenim vrijednostima ne osiguravaju nužno najbolje razdvajanje između podskupina. [4]

**Localized feature selection for clustering**

Bazirano na rješavanju problema selekcije značajki kada imamo skupo podataka gdje prvi podskup značajki dobro definira određeni broj klastera, dok drugi ne. A s druge strane neki drugi podskup značajki dobre definira i ostale klastere, međutim ove prvobitne ne. Selekcijom svih značajki koje pridonose dobrom grupiranju dobiti ćemo podskup koji nemože ukupno dobro grupirati skupo podataka u sve klastere međutim podjelom podskupa na više njih može. Zbog toga je potreban lokalizirana selekcija značajki koja će identificirati više distinktnih podskupova značajki kao rješenje.

Metoda je opisana u članku [5] gdje se koristi kriterij separacije CRIT kao trag umnoška matrice Sw,i koja mjeri koliko su raspršeni uzorci od srednje vrijednosti klastera i te matrice Sb koja mjeri koliko su raspršene srednje vrijednosti klastera od srednje vrijednosti skupa podataka.

CRIT ima svojstvo da je invarijantano prema bilo kojoj nesingularnoj linearnoj transformaciji ako je Sw nesingularna matrica.

CRIT monotono raste s dimenzijama tj. povećanjem broja značajki. Zbog toga je potrebna normalizacija CRIT funkcije u vezi s dimenzijama. Normalizacija se radi unakrsnom projekcijom na pojedinačne klastere. Uvođenjem kazne za pripadanje točaka više klasterima ili nijednom opet prilagođuje CRIT.

Završno algoritam je:

1. Podaci se prvo grupiraju na temelju svih dostupne značajke.
2. Zatim, za svaki klaster, algoritam određuje postoji li suvišna ili bučna značajka
3. Na temelju prilagođene normalizirane vrijednosti definirane prije.
4. Ako postoji biti će uklonjena.

Gornji proces se ponavlja iterativno na svim klasterima sve dok nema promjena. Zatim vraća se rezultat svakog klastera s svojim podskupom značajki. Ukupni proces vidi se na slici ispod.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Kao strategiju pretraživanja koristi sekvencijalnu unazadnu pretragu.

**Nije mi jasno što znači Si kao podskup značajki za određeni klaster.**

Trag matrice je suma vrijednosti dijagonale matrice.

Provodi se algoritam na određenim umjetnim i prirodnim skupovima podataka u uspoređuje s istim algoritmom ali uz globalno selekciju značajki. Lokalna selekcija značajki puno bolje klasificira podatke.[5]

[1] - [Radial basis function - Wikipedia](https://en.wikipedia.org/wiki/Radial_basis_function)

[2] - [Similarity Matrix - Statistics.com: Data Science, Analytics & Statistics Courses](https://www.statistics.com/glossary/similarity-matrix/)

[3] – KNN graf - [Disease\_gene\_identification\_by\_random\_walk\_on\_mult.pdf](file:///C:\Users\bujak\Desktop\FER\5.%20godina\DIPLOMSKI%20PROJEKT\DILPOMSKI%20RAD\Review%20poslane%20literature\Dodatno%20iz%20poslanih%20tekstova\Disease_gene_identification_by_random_walk_on_mult.pdf)

[4] - [A framework for feature selection in clustering - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930825/)

[5] - [doi:10.1016/j.patrec.2007.08.012 (wayne.edu)](http://jinghua.eng.wayne.edu/publication/PRL-LocalizedFeatureSelection.pdf)

* **Metoda omotača** – određuju kvalitetu značajki ovisno o algoritmu grupiranja skupa podataka nakon provedbe algoritma grupiranja

Indeksi ili mjere valjanosti algoritma grupiranja na skupu podataka dijele se na

1. Interni indeksi – mjeri se koliko su dobro raspoređeni klasteri i točke unutar njih od početnog skupa podataka
2. Eksterni indeksi – uspoređuje se sličnost dvaju skupa klastera, najčešće poznati skup i dobiveni skup klastera
   1. Rand indeks
   2. Jaccard indeks
   3. Varijacija informacije

* **Ugrađena metoda** - određuju kvalitetu značajki ovisno o algoritmu grupiranja skupa podataka gdje je kriterij određivanja kvaliteta integriran unutar algoritma grupiranja
* **Hibridna metoda** – koristi i filter metode i metode omotača za selekciju značajki

# An extensive comparative study of cluster validity indices

Usporedba indeks ili mjera valjanosti algoritma grupiranja.

Indeksi ili mjere valjanosti algoritma grupiranja na skupu podataka dijele se na

1. Interni indeksi – mjeri se koliko su dobro raspoređeni klasteri i točke unutar njih od početnog skupa podataka
2. Eksterni indeksi – uspoređuje se sličnost dvaju skupa klastera, najčešće poznati skup i dobiveni skup klastera
   1. Rand indeks
   2. Jaccard indeks
   3. Varijacija informacije

Svojstva koja utječu na kvalitetu klasteriranja:

1. Broj klastera
2. Dimenzije
3. Preklapanje klastera
4. Gustoća klastera
5. Razina buke

Zaključak je da su najbolje mjere Silhouette, Davies–Bouldin, Calinski–Harabasz, generalized Dunn, COP and SDbw. SDbw i Silhouette imaju održano kvalitetu uvođenjem šuma u skupo podatak. Kod promjene gustoće klastera i preklapanja svima proporcionalno padaju rezultati uspješnosti.[6]

[6] - [Sci-Hub | An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243–256 | 10.1016/j.patcog.2012.07.021](https://sci-hub.se/https:/doi.org/10.1016/j.patcog.2012.07.021)

# Semi-supervised feature selection

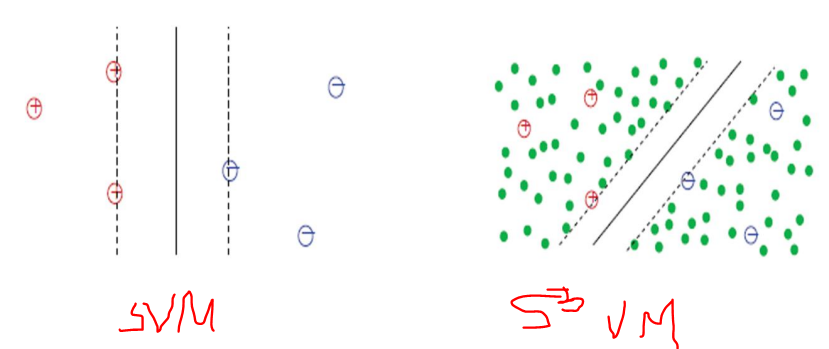
Vrste semi-supervised treniranja klasifikatora:

**Generativni modeli** – Pretpostavljaju p(x,y) = p(y) p(x|y) gdje je p(x|y) distribucija skupa podatka.

**Samotrenirajući** - Kada je manji dio skupa podataka labeliran te se koristi za inicijalno treniranje klasifikatora koji se onda koristi na ne-labeliranom skupu podataka.

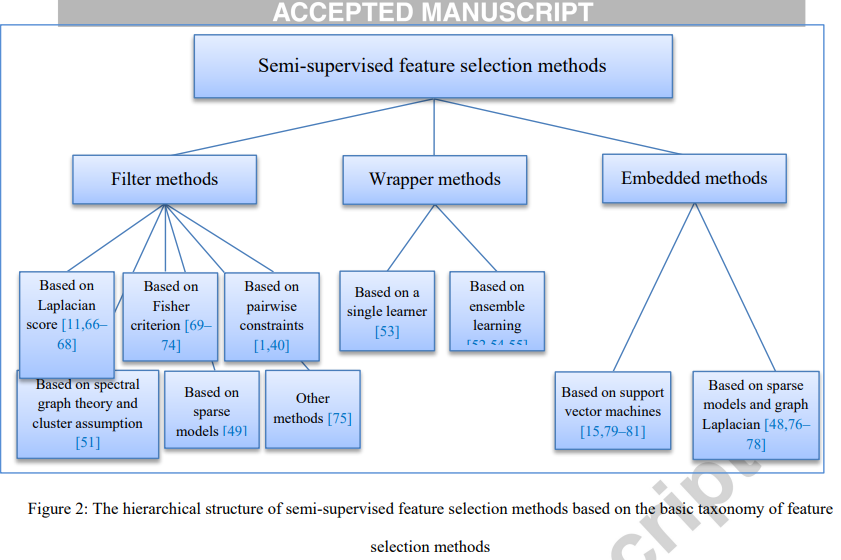
**Ko-trenirajući** – stvaraju se 2 klasifikatora svaki s različitim skupom značajki nad labeliranim skupom podataka. Nakon toga svaki se koristi nad ne-labeliranom skupu podataka s svojim podskupom značajki. Najbolje predikcije svakog klasifikatora na ne-labeliranom skupu podataka se koriste iterativno kod drugog klasifikatora kao labelirani skup podataka.

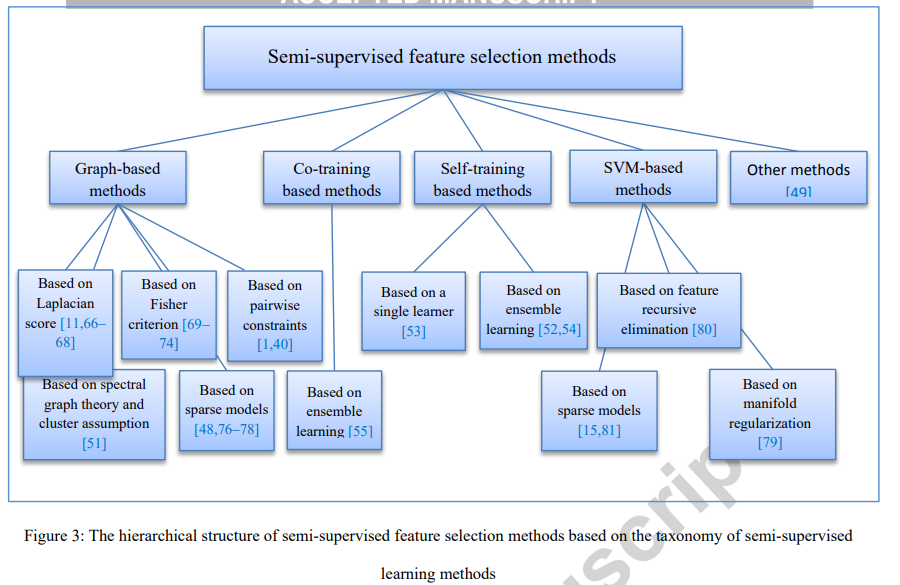
**Semi-supervised metoda potpornih vektora** (S^3VMs)- Koriste nelabelirani i labelirani skup podataka da pronađu linearnu granicu između klastera koja najviše ih razdvaja. Zapravo SVM na labeliranom skupu podataka te ukomponiranje tog u nelabelirani.



**Metode bazirane na grafovima** – Prvo konstruiraju graf pomoću skupa podataka kojem su vrhova nelabelirani i labelirani podaci spojeni vezama koje predstavljaju sličnosti između određenih opservacija.

Podjelu po tipu metoda sem-supervised selekcije značajki vidimo na slici ispod.





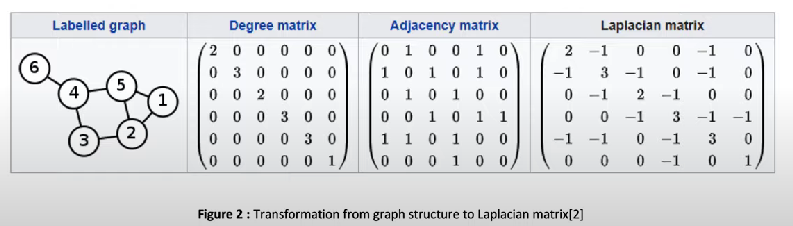
Filter metoda za ne labaleirani skup podataka:

**Variance score** – izbor značajki s najviše varijance po formuli:

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

**Laplacian score** – značajka je dobra ako očuva Lokalnost iz originalnog skupa podatka tj. blizinu točaka.



Za semi-supervised algoritme grade se 2 grafa. Gw graf prikazuje odnose unutar klastera. Čvorovi se spajaju po prijašnjem postupku ako su određene međusobne udaljenosti te ako imaju istu labelu. Gb graf prikazuje odnose između klastera gdje čvorovi se spajaju ako pripadaju suprotnoj klasi.

**Mi nemamo takav skup podataka jer imamo samo jednu sigurnu labelu.**

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Slika na kojoj se prikazuje stol

Opis je automatski generiran