# Izbor metode validacije clustera – Cluster Validity Indeks

11.5.2022.

Metoda validacije clusteriranja treba biti prilagođena algoritmu clusteriranja. Kako se za clusteriranje koristi algoritam baziran na gustoći podataka pozicioniranih kao točke u p-dimenzionalnom prostoru (gdje je p broj dimenzija ili ti značajki) tako je potrebni izabrati mjeru validacije clusteriranja baziranu na gustoći podataka.

Istraživanjem literature metoda validacije clusteriranja ne bazirane na gustoći clustera najčešće se računaju kao omjer minimalne inter-cluster udaljenosti i maksimalne intra-udaljenosti tj. kao izračun razmaka između clustera i kohezije točaka unutar clustera.

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 slučaj gdje su točke unutar klastera najraspršenije.

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 slučaj gdje su dva klastera najbliskija.

Maksimizacijom indeksa znači da su točke unutar klastera bliže raspoređene, a klasteri su međusobno više udaljeni što je optimalni rezultat clusteringa.

Problem kod opisanih metoda što uzimaju i obzir centroide clustera kao središta clustera što je jedino istina kod ovalnih ili hiper-sfernih clustera. Kako podaci u našem skupu podataka nisu hiper-sfernog oblika (barem ne iz vizualizacije podskupa značajki u 3D-u) već proizvoljnog duguljastog oblika tako udaljenost od centroida clustera ne govori ništa od njegovoj kvaliteti. Isto tako korišteni algoritam clusteriranja OPTICS je baziran na gustoći podataka pa tako i onda kvaliteta clustera ovisi o njihov gustoći. Zbog toga je implementirati indeks koji uzima u obzir samu gustoću clustera i između clustera. Navedena vrsta indeksa naziva se indeksi clusteriranja bazirani na gustoći clustera (Density-based Cluster Validity Indices).

Još jedan nedostatak generalnih metoda validacije clusteriranja jest nošenje s outlier podacima. Kako OPTICS algoritam ih relativno kvalitetno detektira potrebno je uvesti metriku kažnjavanja clustering rezultata ako je previše outliera. Metrika je preuzeta iz [1] gdje se završna vrijednost indeksa množi s postotkom outliersa u ukupnom skupu podataka.

Najpoznatiji metode validacije clusteriranja bazirane na gustoći clustera su:

1. SD -
2. S\_Dbw - average scattering for clusters (Scatt)+ between-cluster density (Dbw) [clusterCrit.pdf (r-project.org)](https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf)

Definiran kao zbroj srednje vrijednosti raspršenja clustera te gustoće između clustera.



**Srednja vrijednost raspršenja clustera** definirana je kao srednja vrijednost normi vektora varijanci značajki clustera kroz norma vektora varijanci značajki originalnog skupa podataka.

Vektor varijance čine varijance svakog clustera po formuli (1).

(1),

,

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Prvo se definira radijus hiper-sfere kojom će se ograničiti područje računanja. Radijus hiper-sfere je označen oznakom sigma te iznosi kvadratnom korijenu zbroja norme vektora varijance. Vektor varijance čine varijance svakog clustera po formuli (1).

Slika na kojoj se prikazuje tekst, sat

Opis je automatski generiran

Slijedeće računa se kao relativna gustoća točke ovisne o paru clustera **k** i **k'** koja je jednaka broju točaka unutar ta 2 clustera čija udaljenost do točke je manja od vrijednosti sigma. Geometrijski to je kao uzimanje određene hiper-sfere sa središtem u danoj točci i radijusom sigma te brojanjem svih točaka 2 izabrana clustera unutar definirane hiper-sfere.

Za svaki par clustera dobiva se međusobna gustoća po formuli ispod gdje određuje baricentar (centroid) clustera, a određuje točku na pola puta između 2 clustera (između centroida).

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

**Gustoća između clustera G** je definirana kao srednja vrijednost svih gustoće svih parova clustera prikazana slikom ispod gdje je K ukupni broj clustera.

Slika na kojoj se prikazuje tekst, sat

Opis je automatski generiran

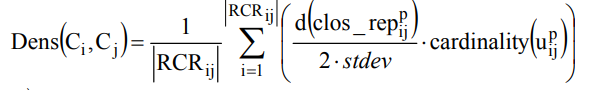
1. CDbw - Composed Density between and within clusters [CDbw: Clustering results’ validity checking based on density and multi-representatives (psu.edu)](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1060.6840&rep=rep1&type=pdf)

Postiže validiranje ne samo hiper-sferičnih clustera zbog korištenja više reprezentativnih točaka za svaki cluster. Ovakav pristup poboljšava geometrijsku efikasnost.

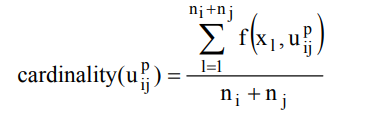
Svaki cluster je reprezentiran sa skupom od **r** točaka gdje je je **j**-ta točka clustera **i**. Skup reprezentativnih točaka clustera su dobro raspršene točke unutar clustera koje trebaju uhvatiti geometriju clustera. Za izbor točaka koristi se metodologija najdalji-prvi. Za prvu reprezentativnu točku clustera uzima se najdalja točka od **centroida** clustera. Svaka slijedeće točka bira se kao najudaljenija točka culustera od prijašnje izabrane točke. Time se točkama definira geometrija granica clustera.

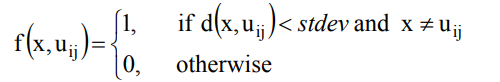
Skup najbližih reprezentativnih točaka **CRij** clustera Ci i Cj definiran je kao skup parova točka iz različitih clustera koje su si najbliže. Respektivni skup najbližih reprezentativnih točaka **RCRij** za clustere Ci i Cj je skup parova točka iz različitih clustera koje su si najbliže reprezentativne točke po prijašnjoj definiciji.

Područje između clustera je definirano kao područje između respektivni skup najbližih reprezentativnih točaka. Gustoća između clustera Ci i Cj definirana je kao



Gdje je sdev srednja standardna devijacija navedenih clustera, d je euklidska udaljenost između para točaka koji su dio respektivnog skupa najbližih reprezentativnih točaka. Cardinality predstavlja kardinalnost između clustera kao





Gdje je x sve točke 2 izabrana clustera te ni i nj njihove veličine. P je p-ti par točaka iz skupa RCRij. Kardinalnost opisuje srednji broj točaka u clusterima Ci i CJ koji pripadaju sujedstvu uij koje hiper-sfera radijusa stdev. d(…)/(2\*stdev) opisuje postotak točaka koje pripadaju susjedstvu uij.

Ukupna Inter-cluster gustoća definirana je kao srednja vrijednost maksimalnih gustoća između para klastera za svaki klaster.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Nadalje seperacija clustera određene je kao odgovarajući broj točaka po jedinici prostora između najbližih clustera.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

1. DBCV - Density-Based Clustering Validation [DBCV.pdf (lmu.de)](https://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf)

Indeks procjenjuje kvalitetu grupiranja na temelju veze relativne gustoće između parova clustera. DBCV indeks je formuliran na temelju nove funkcije gustoće koja se koristi za izračunavanje gustoće clustera i za procjenu intra- i inter- gustoće povezanosti rezultata clusteriranja.

Određuje novu formulu za računanje core-distance točke, validity indexa te nošenje s šumom kod rezultata clusteriranja. Cilj je pronaći područje najmanje gustoće unutar clustera te najveće gustoće izvan clustera.

Mane ostalih cluster validity indexa:

1. S\_Dbw i SD računaju se pomoću centroida clustera što pretpostavlja hiper-sferni oblik clusterea čime onda loše procjenjuju uspjenšnost clusteriranja u arbitrarne oblike.
2. CDbw – računa se uzimajući reprezentativne točke clustera međutim mora se azdati parametar broja reprezentativnih točaka cluster koji jako ovisan o skupu podataka za koji se koristi (veličina, gustoća, oblik). Zbog toga jedinstven broj reprezentativnih točaka za svaki cluster se ne čini kao dobra metodologija pogotovo što izbor reprezentativnih točaka jako utječe na rezultat procjene clusteriranja.

Pojmovi: Računaju se posebno za svaki cluster i točke unutar clustera

1. All points core distance – inverz gustoće svake točke u odnosu na gustoću ostalih točaka unutar clustera
2. Mutual reachability distance od točka i i j (MRD) – maksimum od core distance(i), core distance(j) te d(i,j) za sve parove točka unutar clustera
3. MSTmrd – Minimum spanning tree od grafa gdje su sve točke spojene s reachability distance-om
4. Density sparness clustera (DSC) – maksimum reachability distance od MSTmrd internih poveznica, tj. područje s najmanje gustoće unutar clustera.
5. Density seperation of pair of clusters (DSPC)– minimum MRD između točaka clustera i točaka drugog clustera tj. maksimalna gustoća između 2 clustera. Točke koje se uzimaju u obzir su interni čvorovi MSTmrd svakog clustera.
6. Interni čvorovi i poveznice MST grafa neka budu sve ,osim onih na kraju koji su povezani samo s jednom točkom.
7. Validativnost clustera Ci

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

1. DBCV indeks – poprima vrijednosti između -1 i 1
2. Slika na kojoj se prikazuje tekst, sat

   Opis je automatski generiran

**Je li formula za Core distance računa KNN od točaka unutar clustera ili svih točaka skupa podataka?**

1. CVDD - An internal validity index based on density-involved distance

[An Internal Validity Index Based on Density-Involved Distance | Papers With Code](https://paperswithcode.com/paper/an-internal-validity-index-based-on-density)

Matlab code za CVDD: [hulianyu/CVDD: An Internal Validity Index Based on Density-Involved Distance (github.com)](https://github.com/hulianyu/CVDD)

[1] DBCV - [DBCV.pdf (lmu.de)](https://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf)

Izračun OPTICS algoritma s blokovskim ulazom skupa podataka što će biti potrebno kao real-time implementacija. Analiziram funkciju optics\_merging napisanu u Matlabu. Na ulazu dobiva se zadani broj točaka *crosPointNum* koji se prenosi iz prijašnjeg clusteriranja.

**PROMJENI algoritam da uključuje računanje udaljenosti na temelju svih dimenzija!**

**12.5.2022.**

Radio implementaciju DBCV indexa u Matlabu baziranu na radu [DBCV.pdf (lmu.de)](https://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf).

Pronašao python implementaciju [christopherjenness/DBCV: Python implementation of Density-Based Clustering Validation (github.com)](https://github.com/christopherjenness/DBCV) istog algoritma pa mijenjao kod bazirano na njemu jer nije previše bilo jasno iz originalnog članka.

Rađenje po python implementaciji zahtjeva previše vremena, potrebno refaktoriranje!

**13.5.2022.**

Neko kaže za implementaciju DBCV - [outliers - which metrics are suitable for density-based clustering validation? - Cross Validated (stackexchange.com)](https://stats.stackexchange.com/questions/448988/which-metrics-are-suitable-for-density-based-clustering-validation)

Našao rad koji opisuje korištenje DBCV u matalabu! [1811.07615.pdf (arxiv.org)](https://arxiv.org/pdf/1811.07615.pdf)

Review clustering algoritma koji ima opis DBCV-a  [(arxiv.org)](https://arxiv.org/ftp/arxiv/papers/2106/2106.12792.pdf) te je dao

stranicu koja bira najbolji algoritam ML i metriku validacije za tvoj dataset [My Github Pages | jhillenbrand.github.io](https://jhillenbrand.github.io/cluster_table.html)

Pronašao novi algiroritam density based clusteriranja HDBSCAN –

Originilalni članak - [Sci-Hub | Density-Based Clustering Based on Hierarchical Density Estimates. Lecture Notes in Computer Science, 160–172 | 10.1007/978-3-642-37456-2\_14](https://sci-hub.se/10.1007/978-3-642-37456-2_14)

HDBSCAN za semi-supervised opis - [Sci-Hub | Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data, 10(1), 1–51 | 10.1145/2733381](https://sci-hub.se/10.1145/2733381)

Python code - [scikit-learn-contrib/hdbscan: A high performance implementation of HDBSCAN clustering. (github.com)](https://github.com/scikit-learn-contrib/hdbscan)

Matlab code: [Jorsorokin/HDBSCAN: Hierarchical density based clustering for applications with noise (github.com)](https://github.com/Jorsorokin/HDBSCAN)

Search paper with code: [| Papers With Code](https://paperswithcode.com/task/task)

**14.5.2022.**

Za provjeravanje validnosti DBCV indexa potrebno napraviti korelacijsku usporedbu s eksternalnim indeksom. Eksternali indeksi su mjere koje provjeravaju kvalitetu clusteringa poznavajući ispravnu klasifikaciju podataka u skupu podataka. Glavni eksternalni indeks korišten u literaturi je Adjusted Rand indeks ili ARI. Opis ARI-a nalazi se na [Sci-Hub | Comparing partitions. Journal of Classification, 2(1), 193–218 | 10.1007/bf01908075](https://sci-hub.se/https:/doi.org/10.1007/BF01908075) te Matlab kod na [cmccomb/rand\_index: Computes the rand index (or modified rand index) to describe the agreement between two partitions. (github.com)](https://github.com/cmccomb/rand_index/).

Formula za ARI se je napisana ispod. Gdje je Si i Sj oznaka za skup točaka u clusteru i i j.

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Provjera rand indeks implementacije u matlabu na 8 labeliranih skupova podataka. Za svaki skup podataka pomoću OPTICS algoritma izvodi se clusteriranje s iteriranjem kombinacije parametara epsilon i min\_points. Epsilon je iteriran po vrijednostima 0 do 1 s korakom 0.01 , dok je min\_points iteriran od 5 do 20 s korakom 1. Za svaku kombinaciju epsilon i min\_points parametara dobiveni clusteriran rezultat OPTICS metode uspoređuje se s postojećim labelama skupa podataka pomoću ARI-a. Isto tako za svaku kombinaciju epsilon i min\_points parametara dobiveni clusteriran rezultat OPTICS metode validira se DBCV indeksom.

Kako bi lakše analizirali rezultate za svaku kombinaciju epsilon i min\_points parametra mogu se parametri kombinirati u jedan koji će biti minimalna gustoća potrebna da točka unutar skupa podataka postane dio clustera. Navedena minimalna gustoća će se računati kao površina kruga (za 2D skup podataka, inače volumen hiper-sfere) radijusa epsilon podiljena s min\_points broj točaka. Što znači da je manja minimalna gustoća potrebna da točka unutar skupa podataka postane dio clustera kada je epsilon veći te kada je min\_points manji. Navedeni parametar minimalne gustoće imenovati će se min\_ density\_needed.

Na slici ispod lijevo može se vidjeti 2D prikaz prvog skupa podataka, dok ispod desno labelirani prvi skup podataka gdje je svaki cluster obojan drugom bojom, dok su crveno obojani outlieri.

Za prvi skup podataka ARI daje najbolje rezultate (0.9947) za min\_density\_needed između

**16.5.2022.**

Probani svi indeksi i svi mogućim oblicima na 10 sintetičkih skupova podataka te na 10 realnih skupova podataka.