KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

INTELEKTIKOS PAGRINDAI (P176B101)

Komandinio darbo ataskaita

Atliko:
IFF-1/4 gr. studentas
Mildaras Karvelis
IFF-1/7 gr. studentas
Mantas Vansauskas

2024 m. gegužės 29 d.

Priėmė:
Lekt. Dr. Audrius Nečiūnas

KAUNAS, 2024

Turinys

Projekto aprašymas

Reikalavimai:

- Uždavinys: neprižiūrimojo mašininio mokymosi algoritmų realizacija pasirinktam duomenų rinkiniui
- Komandą sudaro 2 žmonės
- Darbo pateikimo deadline: gegužės 31 diena.
- Ataskaita + programinis kodas

Ką reikia atlikti: Realizuoti 2 neprižiūrimo mašininio mokymosi algoritmus (tarkime *K*-vidurkių ir SOM) ir palyginti gautus rezultatus. Kadangi nėra kaip patikrinti modelio tikslumą (duomenų rinkinyje nėra išvesties) tai turi būti atliktas klasterių vertinimas (pvz., silueto koeficientas). Svarbu atlikti kuo įvairesnius eksperimentus ir pakomentuoti rezultatus

Ataskaitoje reikia pateikti:

- Duomenų rinkinį ir jo aprašymą (gali būti jau naudotas rinkinys (be išvesties tik), arba naujas)
- Metodų, kuriuos planuojate naudoti trumpas aprašymas
- Atstumo metrikos (pvz., *Euklido* arba kita).
- Eksperimentus, kurie apima:
- 1 algoritmas tarkime *SOM*: keisti klasterių skaičių (pvz., k=3, 4, 5) ir pateikti atsakymus grafiškai. Rezultatus pakomentuoti.
- 2 algoritmas tarkime *K-vidurkių*: keisti klasterių skaičių (pvz., k=3, 4, 5) ir pateikti atsakymus grafiškai. Taip pat atlikti eksperimentus pagal skirtingus atributus. Pvz., turite brangakmenių duomenų rinkinį, tai X=aukštis, Y=plotis; X=skaidrumas, Y=svoris;... .
  Paskaičiuoti *inerciją* ir *Siluteto koeficiento*, kurių atsakymus pateikti ir grafiškai, ir skaitinėmis reikšmėmis.
- Palyginti gautus abiejų algoritmų rezultatus ir pateikti savo išvadas, ar duomenų rinkiniui tinkamas klasterizavimo metodas, koks klasterių skaičius geriausias, kokie atributai tinkamiausi ir pan.

Duomenų rinkinys

Duomenų šaltinis: https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering

Atributai:

1. Alcohol (alkoholio kiekis): Alkoholio kiekis

2. Malic Acid (obuolių rūgštis): Ji padidina vyno rūgštingumą ir gali turėti įtakos skoniui.

3. Ash (Pelenai): bendras mineralinių medžiagų kiekis vyne, išmatuotas jį sudeginus.

4. Alcalinity of ash (Pelenų šarmingumas): nustatomi šarminiai pelenų komponentai, rodantys kalio, kalcio ir magnio kiekį.

5. Magnesium (Magnis): pagrindinis mineralas esantis vyne, kuris prisideda prie vyno maistingumo ir gali turėti įtakos skoniui.

6. Total phenols (Bendras fenolų kiekis): visų vyno fenolinių junginių kiekio matas. Jie daro įtaką vyno skoniui, spalvai.

7. Flavanoids (Flavanoidai): fenolinių junginių rūšis, turinti įtakos spalvai ir skoniui, suteikianti kartumo ir aitrumo.

8. Nonflavanoid phenols (Neflavanoidiniai fenoliai): fenolinių junginių klasė, kuri gali turėti kitokią įtaką vyno spalvai ir skoniui nei flavanoidai.

9. Proanthocyanins (Proantocianinai): šie taninų pogrupio junginiai lemia vyno kartuma, aitrumą ir spalvos stabilumą.

10. Color intensity (Spalvos intenstvymas): šis rodiklis parodo vyno spalvos gilumą ir sodrumą, kuriam įtakos gali turėti vynuogių veislė, vyno gamybos būdai ir brandinimo procesas.

11. Hue (Atspalvis): jis parodo tikrąjį vyno spalvos atspalvį, jis suteikia informacijos apie vyno amžių ir naudotą vynuogių veislę.

12. OD280/OD315 of diluted wines (Praskiestų vynų optinis tankis): Jis parodo vyno fenolinių medžiagų kiekį, kuris turi įtakos vyno skoniui ir brandinimo galimybėms.

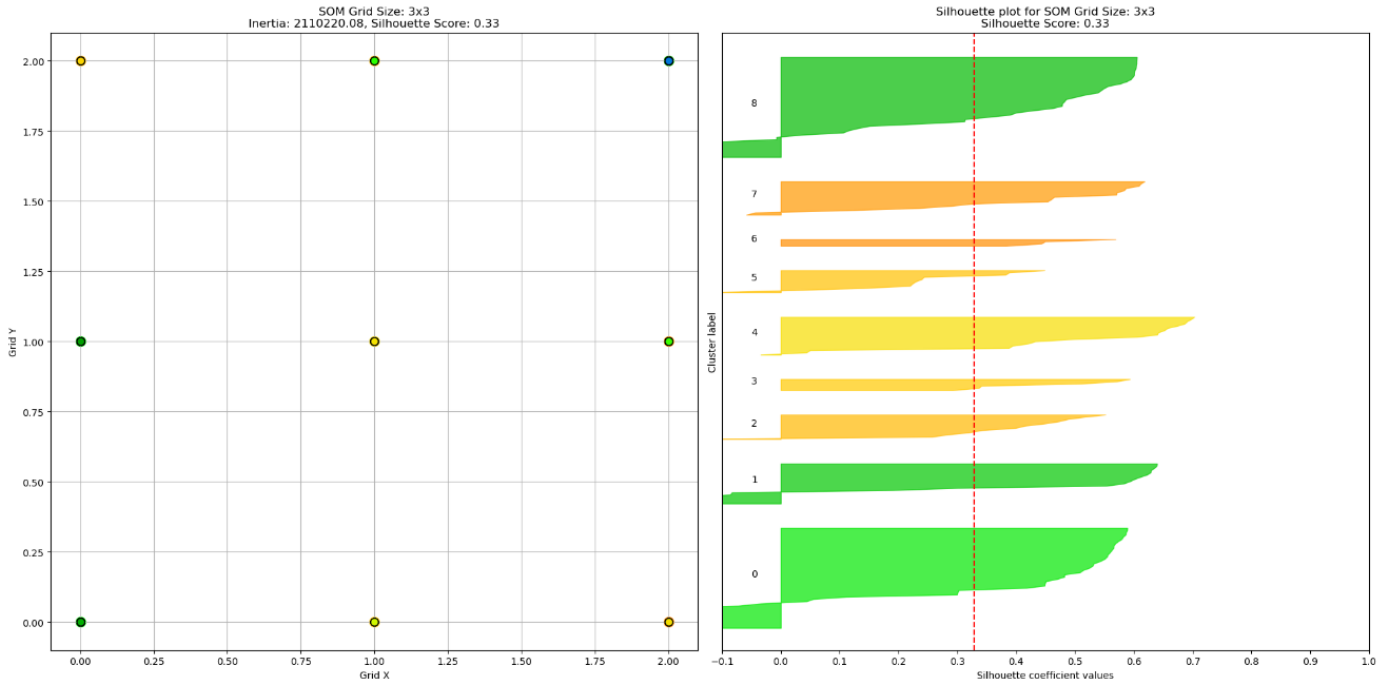
13. Proline (Prolinas): Aminorūgštis, kurios yra vyne ir kuri dažnai naudojama kaip vynuogių prinokimo ir vyno gamybos proceso rodiklis. Didesnis prolino kiekis gali reikšti, kad vynuogės yra labiau subrendusios ir fermentacijos procesas vyksta kitaip.

SOM - Mildaras

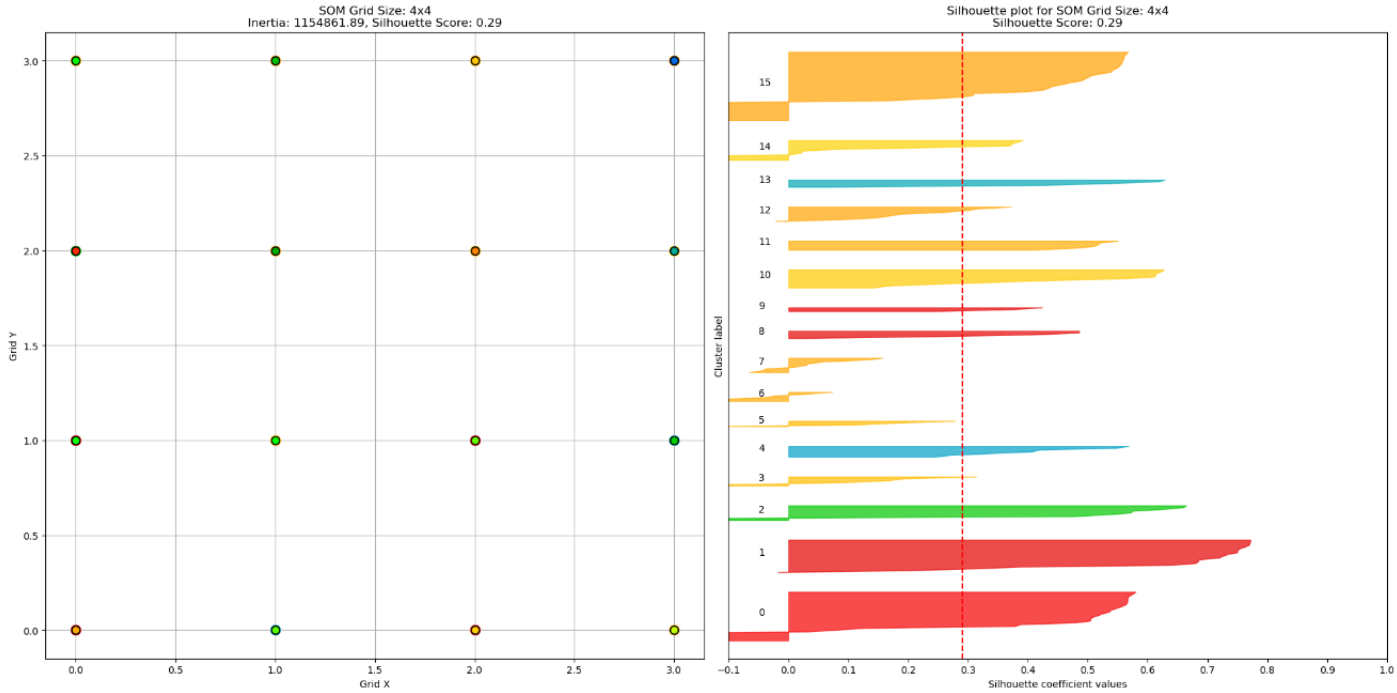
SOM (Self-Organizing Map) yra neprižiūrimas neuroninis tinklas, skirtas duomenų klasifikavimui ir dimensijų mažinimui. Jis transformuoja daugiamačius duomenis į dvimačius arba trimačius tinklo žemėlapio vaizdus, atskirdamas duomenis į grupes pagal jų tarpusavio sąryšius.

SOM grafikai

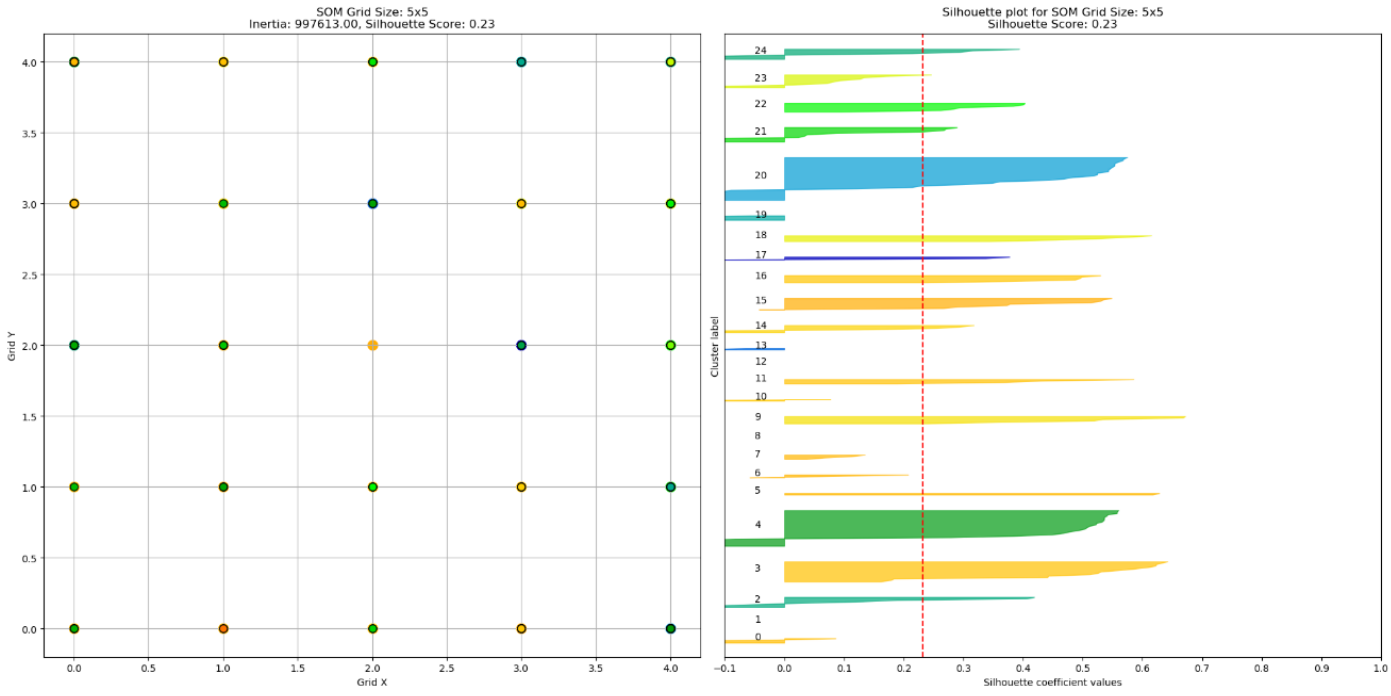
Žemiau pateiktuose grafikuose buvo naudojami 3x3, 4x4 ir 5x5 tinklelio dydžiai.



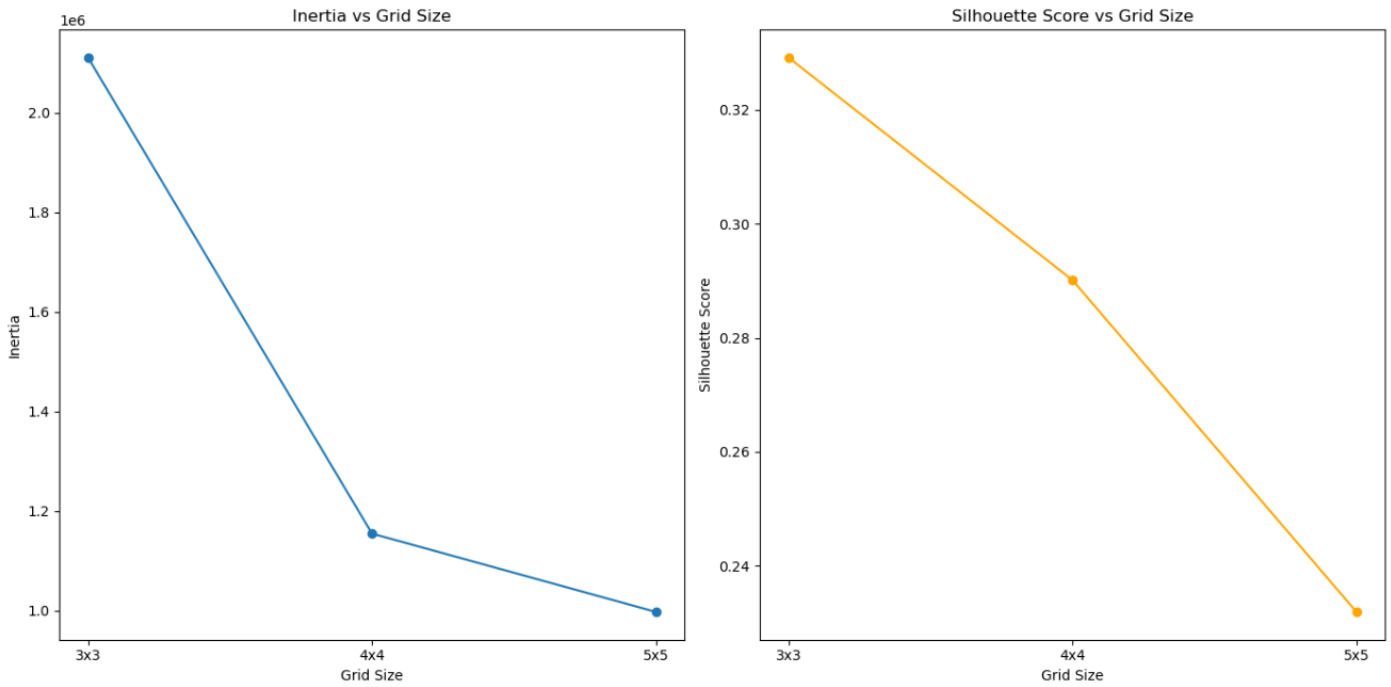
1 grafikas 3x3



2 grafikas 4x4



3 grafikas 5x5



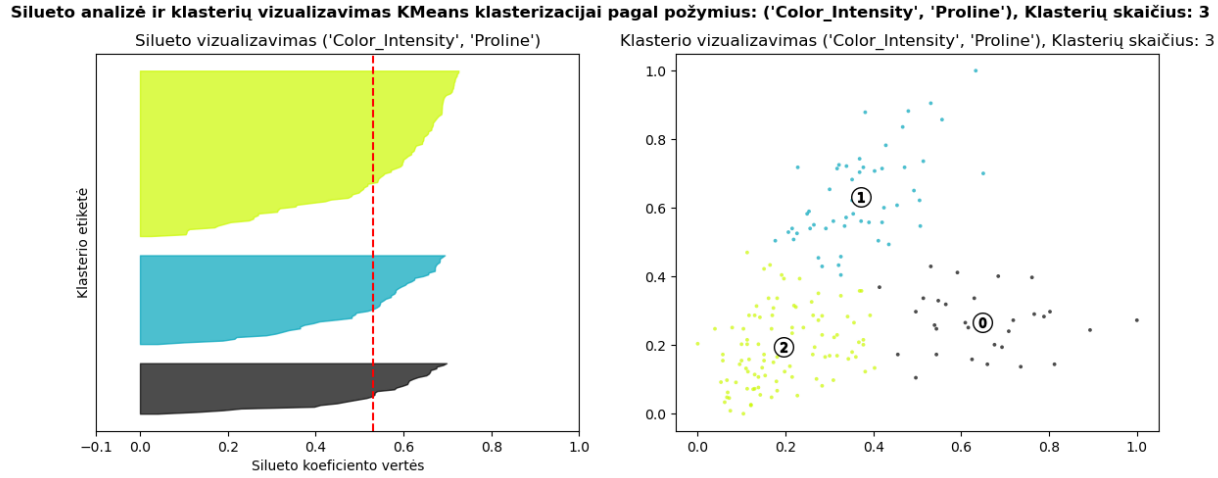
4 grafikas Visų tinklelio dydžių reikšmės

Aukščiau pateiktame grafike (4 grafikas.) matome kaip kinta Inercijos ir Silueto koeficiento reikšmės, didėjant tinklelio dydžiui koeficientai mažėja. Galime teigti, kad SOM algoritmas yra netinkamas šiam duomenų rinkiniui.

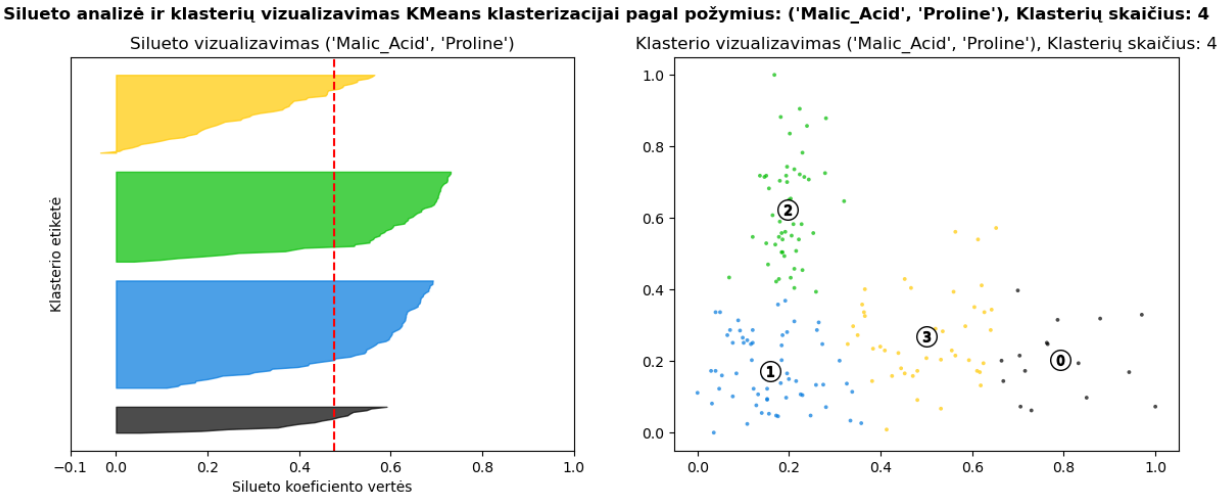
K-Vidurkių - Mantas

K vidurkių algoritmas yra populiarus klasterizavimo metodas, skirtas duomenų taškų grupavimui į K iš anksto nustatytų klasterių. Jis veikia pasirenkant K pradinių centrų, priskiriant kiekvieną tašką artimiausiam centrui ir atnaujinant centrus kaip priskirtų taškų vidurkį. Šie žingsniai kartojami tol, kol centrai stabilizuojasi arba pasiekiamas maksimalus iteracijų skaičius. Algoritmas yra greitas ir lengvai suprantamas.

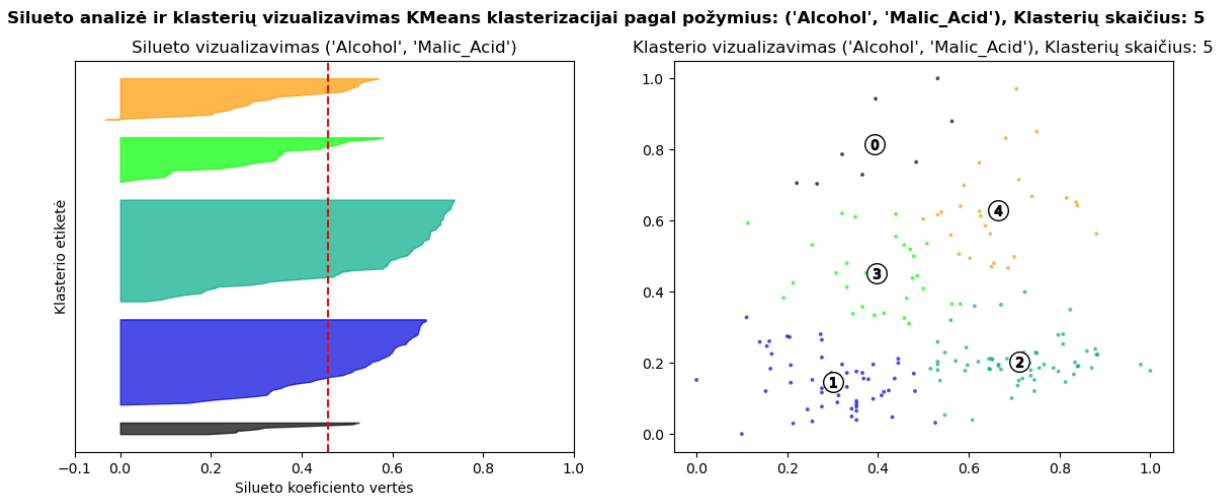
K-Vidurkių grafikai



Geriausius rezultatus esant trims klasteriams gavome, jog ('Color_Intensity', 'Proline') turi didžiausią Silueto koeficientą = 0.531



Geriausius rezultatus esant keturiems klasteriams gavome, jog ('Malic_Acid', 'Proline') turi didžiausią Silueto koeficientą = 0.477



Geriausius rezultatus esant penkiems klasteriams gavome, jog ('Alcohol', 'Malic_Acid') turi didžiausią Silueto koeficientą = 0.458

Rezultatų palyginimas

Lyginant SOM ir K-Vidurkio gautus rezultatus, matome, kad K-Vidurkio reikšmės stipriai pranoksta SOM reikšmes.

Geriausia K-Vidurkių reikšmė yra su 3 klasteriais, Silueto koeficientas = 0.531

Geriausia SOM reikšmė yra esant 3x3 tinklelio dydžiui, Silueto koeficientas = 0.33