Tema: Klasteriranje i analiza obrazaca ekosustava

Kolegij: Strojno učenje za primjene u okolišu

Ishodi učenja

Nakon završetka ove nastavne jedinice studenti će moći:

- Primijeniti tehnike klasteriranja za prepoznavanje obrazaca i grupiranja u skupovima podataka o okolišu.
- Vrednovati kvalitetu klastera primjenom unutarnjih i vanjskih validacijskih metrika.
- Tumačiti ekološko značenje nenadziranih modela.
- Koristiti rezultate klasteriranja kao potporu očuvanju prirode i upravljanju ekosustavima.

Uvod

Klasteriranje je temeljni pristup nenadziranog učenja koji organizira uzorke podataka u skupine na temelju sličnosti, bez potrebe za unaprijed definiranim oznakama. U analitici okoliša posebno je vrijedno jer ekosustavi često pokazuju složene i višedimenzionalne strukture u kojima jasne granice klasa nisu očite. Prepoznavanjem prirodnih grupiranja unutar podataka, klasteriranje podupire zadatke kao što su procjena bioraznolikosti, analiza kemijskih svojstava tla i tumačenje opažanja daljinskih istraživanja, uz istodobno uvažavanje šuma i nepravilnih obrazaca koji su česti u ekološkim mjerenjima [4,6,8]. Nenadzirano klasteriranje primijenjeno na podatke o produktivnosti dobivene daljinskim istraživanjima povezano je s obrascima bioraznolikosti velikih razmjera [2].

K-Means klasteriranje

K-Means dijeli skup podataka na k klastera tako da iterativno dodjeljuje točke najbližem centroidu i ažurira položaje centroida radi minimiziranja varijance unutar klastera. To je jedan od najčešće korištenih algoritama klasteriranja zbog svoje jednostavnosti i računalne učinkovitosti.

Prednosti:

- Brz je i skalabilan na velike skupove podataka.
- Jednostavan je za implementaciju i tumačenje.

Ograničenja:

- Zahtijeva da se broj klastera odredi unaprijed.k
- Osjetljiv je na inicijalizaciju i odstupajuće vrijednosti.
- Najbolje funkcionira s kuglastim klasterima podjednake veličine.

Primjene u okolišu: otkrivanje obrazaca onečišćenja u mjerenjima kvalitete vode [9].

Ključno ograničenje algoritma K-Means jest zahtjev da se broj klastera (k) zada unaprijed. U primjenama u okolišu stvarni broj režima ekosustava rijetko je unaprijed poznat, a neprimjeren izbor vrijednosti k može dovesti do pretjerano pojednostavljenih ili fragmentiranih prikaza ekološke strukture.

Kako bi se taj problem ublažio, za usmjeravanje odabira vrijednosti k često se koriste unutarnje validacijske tehnike. Te metode procjenjuju kvalitetu klasteriranja na temelju kompaktnosti i razdvojenosti klastera, bez oslanjanja na vanjske oznake. Među široko korištenim pristupima su metoda lakta, koja ispituje smanjenje varijance unutar klastera kako se k povećava, i analiza Silhouette, koja procjenjuje koliko se dobro svaka podatkovna točka uklapa u dodijeljeni klaster u usporedbi sa susjednim klasterima.

U praksi konačni odabir vrijednosti k često uravnotežuje rezultate kvantitativne validacije s ekološkom interpretabilnošću, čime se osigurava da prepoznati klasteri predstavljaju smislene režime okoliša, a ne samo matematičke podjele.

Hijerarhijsko klasteriranje

Hijerarhijsko klasteriranje stvara hijerarhiju klastera, često prikazanu dendrogramom, koji otkriva ugniježđene razine sličnosti među uzorcima. Može biti aglomerativno (odozdo prema gore) ili divizivno (odozgo prema dolje). Ova je metoda osobito korisna u ekološkim istraživanjima u kojima je potrebno razumjeti odnose na više prostornih ili organizacijskih razina.

Prednosti:

- Ne zahtijeva unaprijed određivanje broja klastera.
- Stvara dendrograme koji otkrivaju višerazinske ekološke strukture.
- Omogućuje fleksibilne kriterije povezivanja (single, complete, average).

Ograničenja:

- Računalno je zahtjevno za vrlo velike skupove podataka.
- Osjetljivo je na šum i skaliranje varijabli.
- Rezultate može biti teško tumačiti ako klasteri nisu dobro razdvojeni.

Primjene u okolišu: klasteriranje hidroloških slivova na temelju režima protoka i oborina [5].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN grupira točke koje su gusto raspoređene, dok točke u područjima male gustoće označuje kao šum. Za razliku od algoritma K-Means, ne zahtijeva da se broj klastera zada unaprijed i može prepoznati klastere proizvoljnog oblika.

Prednosti:

- Može otkriti klastere nepravilnog oblika.
- Automatski prepoznaje šum/odstupajuće vrijednosti.
- Nije potrebno unaprijed definirati broj klastera.

Ograničenja:

- Učinkovitost snažno ovisi o izboru parametara (epsilon, minPts).
- Teško se nosi s različitim gustoćama klastera.
- Manje je skalabilan od algoritma K-Means za vrlo velike skupove podataka.

Primjene u okolišu: otkrivanje žarišta onečišćenja u mrežama senzora kvalitete zraka ili vode [3, 7].

Metrike vrednovanja klastera

Za razliku od nadzirane klasifikacije, klasteriranju u većini skupova podataka o okolišu nedostaju stvarne referentne oznake. Vrednovanje kvalitete klasteriranja stoga se oslanja na unutarnje ili vanjske validacijske indekse koji kvantificiraju kompaktnost, razdvojenost ili slaganje s grupiranjima koja su definirali stručnjaci.

Silhouette Score

- Mjeri koliko je svaka točka slična vlastitom klasteru u usporedbi s drugim klasterima.
- Vrijednosti su u rasponu od -1 do 1; više vrijednosti upućuju na dobro razdvojene i kohezivne klastere.
- Korisna je za odabir broja klastera k u algoritmu K-Means.

Davies–Bouldin indeks

- Omjer raspršenosti unutar klastera i razdvojenosti između klastera.
- Niže vrijednosti upućuju na bolje klasteriranje.

- Računalno je učinkovit i primjenjiv na velike skupove podataka o okolišu.

Calinski–Harabasz indeks

- Uspoređuje varijancu između klastera s varijancom unutar klastera.
- Više vrijednosti odgovaraju bolje definiranim klasterima.
- Često se koristi u ekološkim primjenama s višedimenzionalnim podacima.

Prilagođeni Randov indeks (ARI)

- Vanjska validacijska metrika koja uspoređuje rezultate klasteriranja s podacima označenima od strane stručnjaka ili s poznatim ekološkim regijama.
- Prilagođen je za slučajnost, pri čemu 1 označuje savršeno slaganje, a 0 slučajnu podjelu.
- Vrijedan je pri validaciji u odnosu na uspostavljene karte ekoregija ili jedinice očuvanja.

Validacija klastera osigurava da otkrivene strukture nisu artefakti algoritma, nego da odražavaju smislene ekološke obrasce. U istraživanjima okoliša kombiniranje unutarnjih indeksa (npr. Silhouette, Davies–Bouldin) s vanjskom validacijom u odnosu na stručne klasifikacije pruža čvrste dokaze za ekološka tumačenja.

Studija slučaja: Prepoznavanje režima ekosustava primjenom nenadziranog klasteriranja podataka o okolišu

Cilj

Ova studija slučaja pokazuje kako se tehnike nenadziranog učenja mogu koristiti za otkrivanje prirodnih grupiranja i ekoloških režima u podacima o okolišu bez unaprijed definiranih oznaka. Cilj je prepoznati regije sa sličnim okolišnim obilježjima i tumačiti te klastere kao različite režime ekosustava ili okoliša. Takvo otkrivanje obrazaca osobito je vrijedno u ekološkoj analizi, gdje su prirodne granice često postupne, nejasne ili nepoznate.

Korištene biblioteke

Studija slučaja implementirana je u Pythonu, uz korištenje široko prihvaćenih biblioteka otvorenog koda za analizu podataka, strojno učenje i vizualizaciju:

- **NumPy**
  Koristi se za numeričke operacije, manipulaciju poljima i slučajno uzorkovanje.

- **Pandas**
  Koristi se za učitavanje, obradu i predobradu tabličnih skupova podataka o okolišu.
- **Scikit-learn (sklearn).**
  Koristi se za:
  - predobradu podataka (StandardScaler),
  - algoritme klasteriranja (K-Means),
  - smanjenje dimenzionalnosti (PCA),
  - unutarnje validacijske metrike klastera (Silhouette Score, Davies–Bouldin indeks).
- **SciPy**
  Koristi se za hijerarhijsko klasteriranje i vizualizaciju dendrograma.
- **Matplotlib**
  Koristi se za vizualizaciju rezultata klasteriranja, dijagrama raspršenja, PCA projekcija i dendrograma.

Ovaj odabir podupire transparentnost i omogućuje jednostavno izvođenje radnog tijeka u standardnim okruženjima kao što su Jupyter Notebook ili Google Colab.

Izvor podataka

Analiza se temelji na skupu podataka Forest Cover Type (Covertype) [1], otvoreno dostupnom skupu podataka koji pruža UCI Machine Learning Repository. Skup podataka sadržava okolišna i topografska mjerenja prikupljena iz šumskih područja u Roosevelt National Forestu u SAD-u. Iako skup podataka uključuje oznaku zemljišnog pokrova, ta se informacija u ovoj studiji slučaja namjerno isključuje kako bi se očuvala nenadzirana narav analize.

Opis podataka

Svako opažanje predstavlja geografsku lokaciju opisanu okolišnim varijablama kao što su:

- Nadmorska visina
- Nagib
- Ekspozicija padine
- Horizontalne i vertikalne udaljenosti do hidroloških obilježja
- Udaljenost do prometnica
- Indeksi zasjenjenja terena

Te varijable zajedno opisuju teren, gradijente vlage i dostupnost, što su ključni pokretači diferencijacije ekosustava.

Predobrada podataka

Za pripremu skupa podataka za klasteriranje primjenjuju se sljedeći koraci:

- Odabir samo kontinuiranih okolišnih i topografskih varijabli
- Uklanjanje oznake tipa šumskog pokrova
- Standardizacija varijabli na srednju vrijednost nula i jediničnu varijancu kako bi se osigurao jednak doprinos metodama klasteriranja temeljenima na udaljenosti

```
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Load dataset (CSV version of Covertype)
df = pd.read_csv("covtype.csv")

# Drop the cover type label (last column)
X = df.iloc[:, :-1]

# Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

pd.DataFrame(X_scaled, columns=X.columns).head()
```

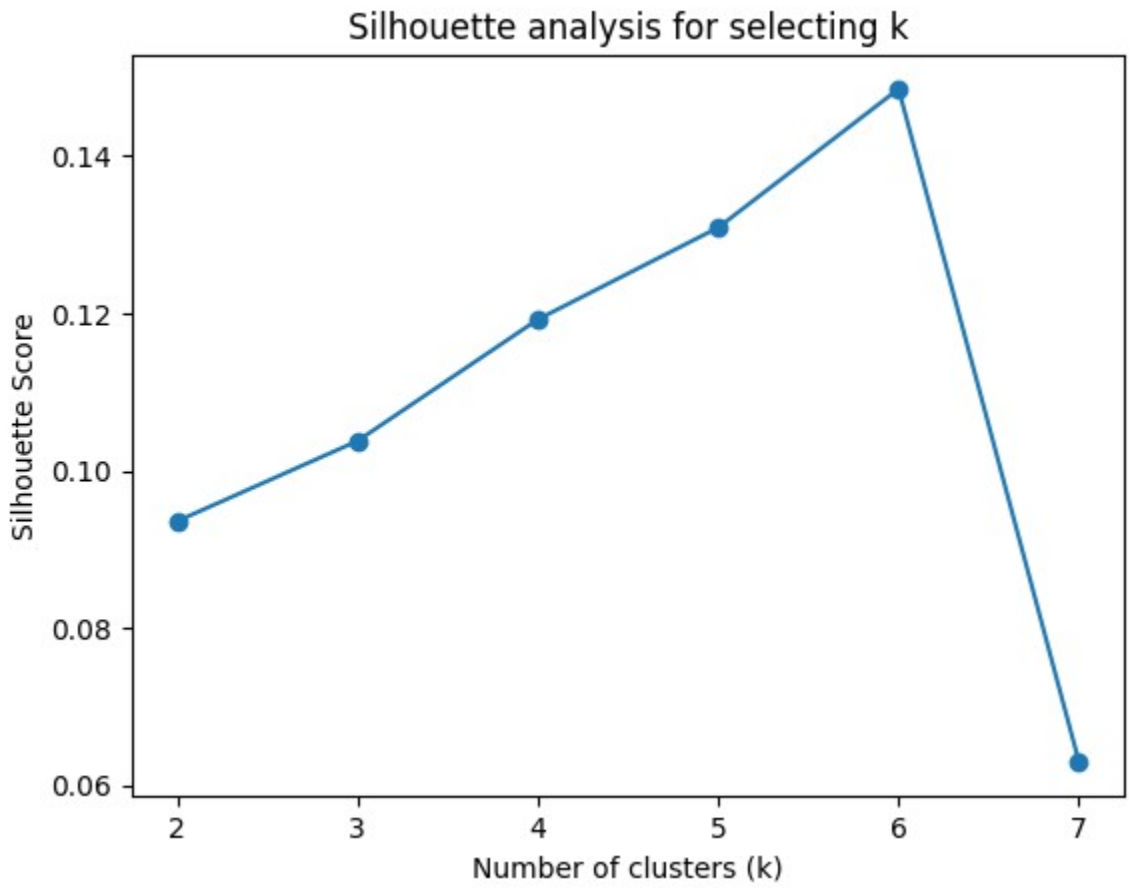
Metode klasteriranja

Primjenjuju se dvije komplementarne tehnike klasteriranja:

K-Means klasteriranje

K-Means se koristi za podjelu skupa podataka na unaprijed definiran broj klastera na temelju okolišne sličnosti. Prije primjene K-Means klasteriranja broj klastera ispitan je primjenom unutarnjih validacijskih tehnika. Analiza Silhouette korištena je za procjenu kvalitete klasteriranja za različite vrijednosti k. Te analize podupiru odabir broja klastera koji uravnotežuje razdvojenost, kompaktnost i ekološku interpretabilnost.

```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt

# Range of k values to test
K_range = range(2, 8)

silhouette_scores = []

for k in K_range:
    print(f"Computing Silhouette Score for k = {k} ...")

    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(X_scaled)

    score = silhouette_score(X_scaled, labels)
    silhouette_scores.append(score)

# Plot results
plt.figure()
plt.plot(K_range, silhouette_scores, marker='o')
plt.xlabel("Number of clusters (k)")
plt.ylabel("Silhouette Score")
plt.title("Silhouette analysis for selecting k (full dataset)")
plt.show()
```



Analiza Silhouette pokazala je više vrijednosti za k = 6, što upućuje na bolju razdvojenost klastera iz čisto statističke perspektive. Međutim, u primjenama u okolišu optimalan broj klastera ne određuje se isključivo unutarnjim validacijskim metrikama. Ekosustavima upravljaju kontinuirani gradijenti, a ne oštro razdvojene klase, pa povećanje broja klastera može fragmentirati ekološki koherentne režime u manje, slabije interpretabilne jedinice.

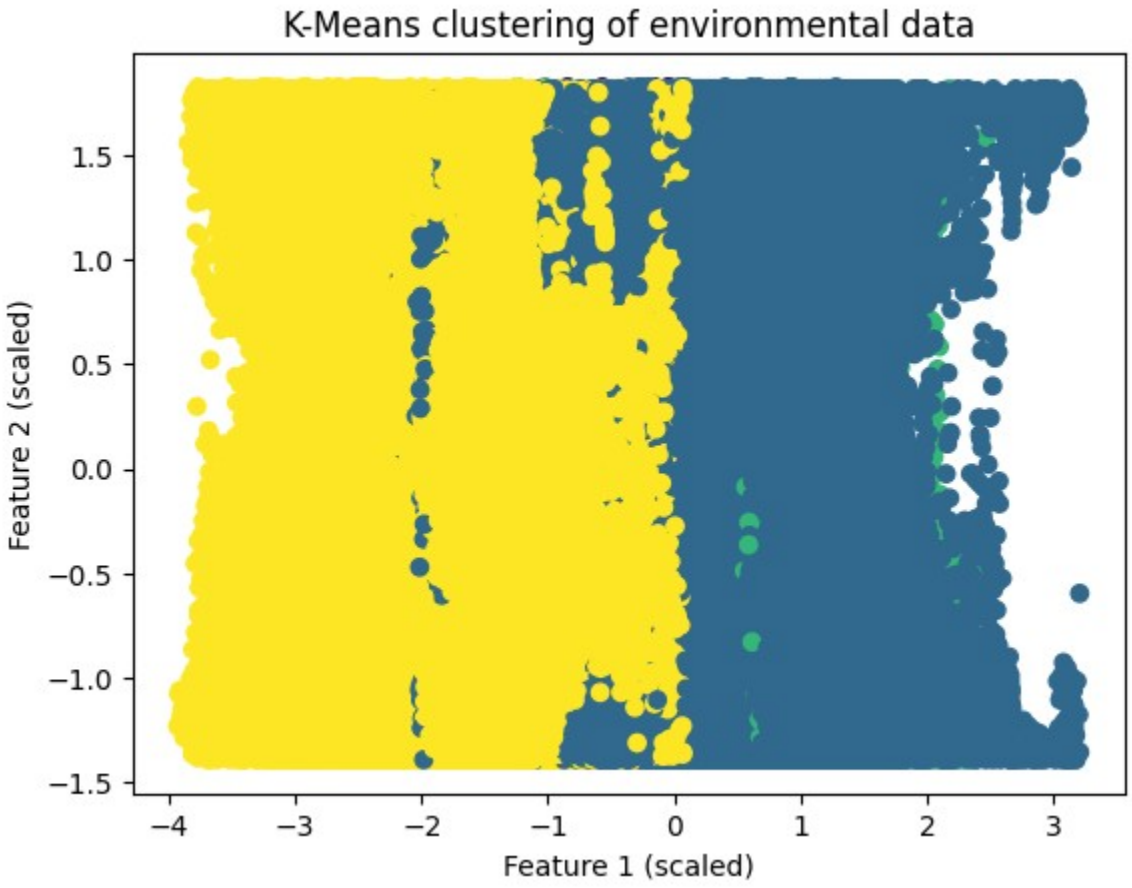
Stoga je odabran manji broj klastera (k = 4) kako bi se uravnotežila statistička kvaliteta klasteriranja s ekološkom interpretabilnošću i pedagoškom jasnoćom. Taj odabir omogućuje prepoznavanje širokih i smislenih režima ekosustava koji se podudaraju s dominantnim okolišnim gradijentima, uz izbjegavanje pretjeranog particioniranja koje bi moglo otežati praktično razumijevanje i primjenu.

```python
from sklearn.cluster import KMeans

# Apply K-Means
kmeans = KMeans(n_clusters=4, random_state=42)
clusters = kmeans.fit_predict(X_scaled)

# Add cluster labels to dataframe
df["Cluster"] = clusters

df[["Cluster"]].value_counts()
```

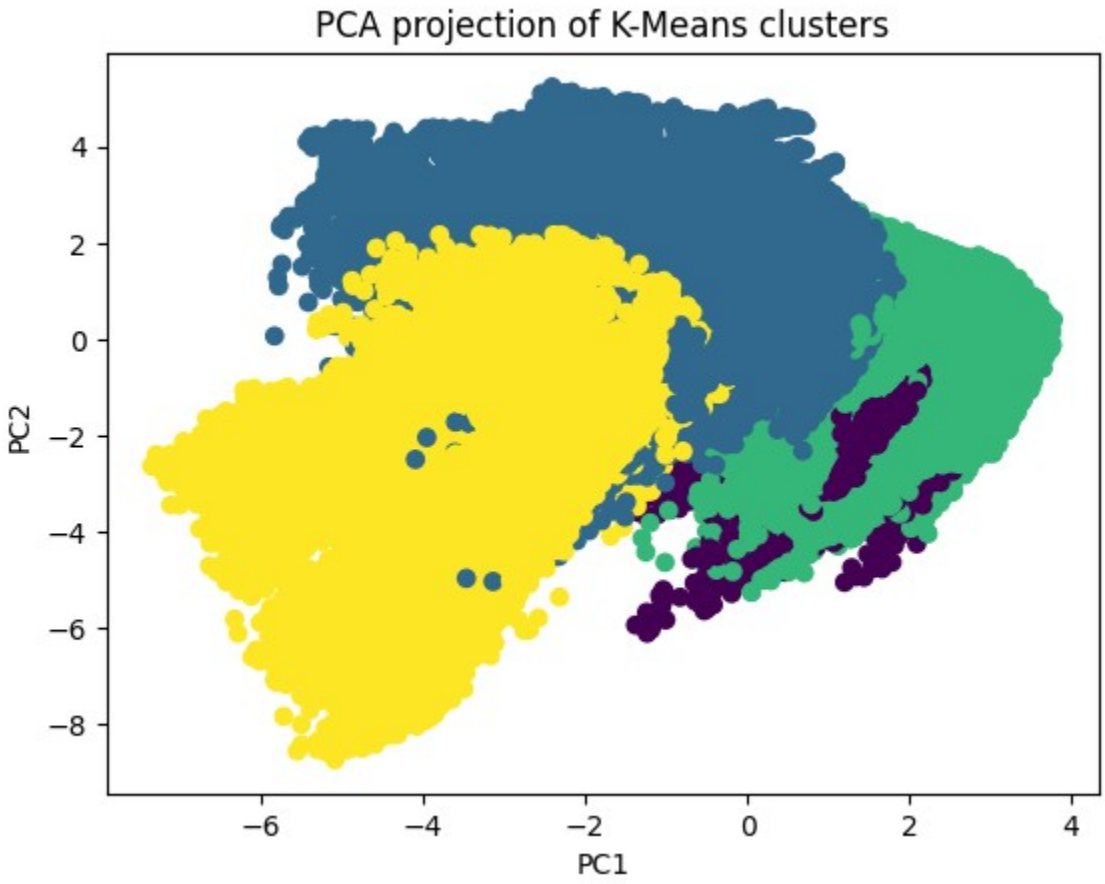
✓ 1.9s

```
Cluster
1          236975
2          228173
3           84547
0           31317
Name: count, dtype: int64
```

```python
import matplotlib.pyplot as plt

plt.figure()
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=clusters)
plt.xlabel("Feature 1 (scaled)")
plt.ylabel("Feature 2 (scaled)")
plt.title("K-Means clustering of environmental data")
plt.show()
```



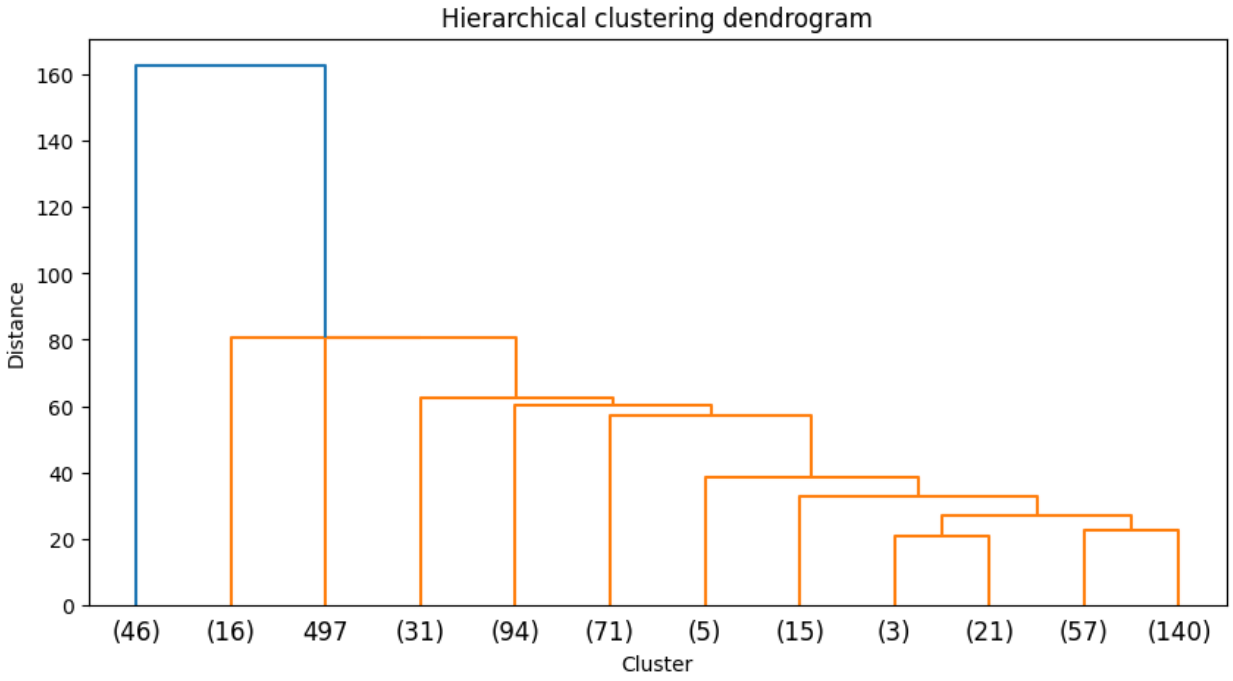

```python
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Apply PCA to reduce to 2 dimensions
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure()
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=clusters)
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("PCA projection of K-Means clusters")
plt.show()
```



Hijerarhijsko klasteriranje

Hijerarhijsko aglomerativno klasteriranje primjenjuje se za istraživanje ugniježđenih odnosa među lokacijama uzorkovanja. Dendrogrami se koriste za vizualizaciju višerazinskih ekoloških struktura i za prikaz načina na koji se lokalni obrasci agregiraju u šira grupiranja ekosustava.

```python
from scipy.cluster.hierarchy import linkage, dendrogram

# Use a subset for visualization
X_subset = X_scaled[:500]

linked = linkage(X_subset, method="ward")

plt.figure(figsize=(10, 5))
dendrogram(linked, truncate_mode="lastp", p=12)
plt.title("Hierarchical clustering dendrogram")
plt.xlabel("Cluster")
plt.ylabel("Distance")
plt.show()
```



Vrednovanje klastera

Kvaliteta klastera procjenjuje se primjenom unutarnjih validacijskih indeksa:

- Silhouette Score
- Davies–Bouldin indeks

Te metrike pomažu osigurati da su prepoznati klasteri kompaktni, dobro razdvojeni i da nisu artefakti slučajne podjele.

```
from sklearn.metrics import silhouette_score, davies_bouldin_score

sil_score = silhouette_score(X_scaled, clusters)
db_score = davies_bouldin_score(X_scaled, clusters)

sil_score, db_score
```

Izračun unutarnjih validacijskih metrika kao što su Silhouette Score i Davies–Bouldin indeks može biti računalno zahtjevan za velike skupove podataka o okolišu jer se te metrike oslanjaju na izračune parnih udaljenosti. S obzirom na veličinu skupa podataka korištenog u ovoj studiji slučaja, procjena kvalitete klasteriranja stoga je provedena na reprezentativnom slučajnom podskupu podataka. Takav pristup čuva temeljnu strukturu podataka, a istodobno osigurava računalnu izvedivost, te je uobičajena praksa u analizama klasteriranja velikih razmjera. Dobivene metrike pružaju pouzdane usporedne uvide u kvalitetu klasteriranja bez potrebe za iscrpnim računanjem nad cijelim skupom podataka.

```
import numpy as np
from sklearn.metrics import silhouette_score, davies_bouldin_score

# Subsample for faster evaluation
np.random.seed(42)
sample_idx = np.random.choice(len(X_scaled), size=5000, replace=False)

X_sample = X_scaled[sample_idx]
clusters_sample = clusters[sample_idx]

sil_score = silhouette_score(X_sample, clusters_sample)
db_score = davies_bouldin_score(X_sample, clusters_sample)

sil_score, db_score
```

Izračun unutarnjih validacijskih metrika kao što su Silhouette Score i Davies–Bouldin indeks može biti računalno zahtjevan za velike skupove podataka o okolišu jer se te metrike oslanjaju na izračune parnih udaljenosti. S obzirom na veličinu skupa podataka korištenog u ovoj studiji slučaja, procjena kvalitete klasteriranja stoga je provedena na reprezentativnom slučajnom

podskupu podataka. Takav pristup čuva temeljnu strukturu podataka, a istodobno osigurava računalnu izvedivost, te je uobičajena praksa u analizama klasteriranja velikih razmjera. Dobivene metrike pružaju pouzdane usporedne uvide u kvalitetu klasteriranja bez potrebe za iscrpnim računanjem nad cijelim skupom podataka.

```
import numpy as np
from sklearn.metrics import silhouette_score, davies_bouldin_score

# Subsample for faster evaluation
np.random.seed(42)
sample_idx = np.random.choice(len(X_scaled), size=5000, replace=False)

X_sample = X_scaled[sample_idx]
clusters_sample = clusters[sample_idx]

sil_score = silhouette_score(X_sample, clusters_sample)
db_score = davies_bouldin_score(X_sample, clusters_sample)

sil_score, db_score
```

Tumačenje i ekološko značenje

Dobiveni klasteri mogu se tumačiti kao različiti režimi ekosustava, primjerice:

- Šumska okruženja niske nadmorske visine nasuprot šumskim okruženjima velike nadmorske visine
- Vlažni nasuprot suhim uvjetima terena
- Dostupne nasuprot udaljenim planinskim regijama

Usporedbom klastera s poznatim okolišnim gradijentima studenti mogu povezati svaku skupinu s karakterističnim ekološkim uvjetima, a ne s unaprijed definiranim klasama zemljišnog pokrova.

Važnost za upravljanje ekosustavima

Otkriveni klasteri pružaju objektivnu osnovu za:

- Klasifikaciju i kartiranje ekosustava
- Prepoznavanje okolišno sličnih regija za planiranje očuvanja

- Potporu odlukama o upravljanju korištenjem zemljišta u područjima s nedostatkom podataka ili slabom kartiranošću

Ova studija slučaja pokazuje kako klasteriranje nadopunjuje stručno znanje otkrivanjem ekoloških struktura vođenih podacima koje se ne moraju podudarati s administrativnim granicama.

Zadaci za vježbu

- Preuzmite skup podataka Forest Cover Type (Covertype) iz UCI Machine Learning Repositoryja [1].
- Istražite skup podataka i provedite početnu deskriptivnu analizu okolišnih i topografskih varijabli.
- Odaberite podskup relevantnih kontinuiranih varijabli (npr. nadmorska visina, nagib, udaljenosti do hidroloških obilježja) i obrazložite odabir.
- Standardizirajte odabrane varijable i ispitajte kako skaliranje utječe na klasteriranje temeljeno na udaljenosti.
- Primijenite unutarnje validacijske tehnike (metodu lakta i analizu Silhouette) za istraživanje različitih vrijednosti broja klastera (k).
- Primijenite K-Means klasteriranje uz odabranu vrijednost k i analizirajte dobivenu strukturu klastera.
- Provedite hijerarhijsko klasteriranje na smanjenom podskupu podataka i vizualizirajte višerazinske odnose pomoću dendrograma.
- Usporedite rezultate klasteriranja dobivene metodama K-Means i hijerarhijskim metodama, ističući sličnosti i razlike.
- Tumačite klastere s obzirom na dominantne okolišne gradijente, a ne na unaprijed definirane klase zemljišnog pokrova.
- Razmotrite ograničenja klasteriranja za analizu ekosustava i raspravite kako bi rezultati mogli nadopuniti stručnu klasifikaciju ekosustava i odluke o upravljanju.

Provjera znanja

Koja je glavna svrha primjene tehnika klasteriranja u analizi obrazaca ekosustava?

A. Predvidjeti buduće okolišne varijable
B. Dodijeliti uzorke unaprijed definiranim klasama rizika
C. Prepoznati prirodna grupiranja i obrasce u neoznačenim podacima o okolišu
D. Smanjiti dimenzionalnost isključivo radi vizualizacije

Koje je ograničenje najkarakterističnije za algoritam klasteriranja K-Means u skupovima podataka o okolišu?

A. Nemogućnost rada s velikim skupovima podataka
B. Zahtjev da se broj klastera (k) unaprijed definira
C. Nemogućnost rada s numeričkim varijablama
D. Nemogućnost otkrivanja kompaktnih klastera

Koja je metoda klasteriranja najprikladnija za otkrivanje žarišta onečišćenja nepravilnog oblika uz prepoznavanje šuma?

A. K-Means
B. Hijerarhijsko klasteriranje
C. DBSCAN
D. Logistička regresija

Zašto se unutarnje validacijske metrike kao što je Silhouette Score često koriste u analizi klasteriranja?

A. Zato što klasteriranje uvijek ima stvarne referentne oznake
B. Za predviđanje budućih promjena ekosustava
C. Za vrednovanje kompaktnosti i razdvojenosti klastera bez označenih podataka
D. Za optimizaciju nadziranih klasifikatora

Kako rezultati klasteriranja mogu poduprijeti odluke o upravljanju ekosustavima i očuvanju prirode?

A. Zamjenom ekoloških terenskih opažanja
B. Pružanjem točnih prognoza promjene ekosustava
C. Otkrivanjem funkcionalnih skupina i ekoloških gradijenata relevantnih za upravljanje
D. Uklanjanjem nesigurnosti u podacima o okolišu

Dodatni izvori

Machine Learning - A First Course for Engineers and Scientists, Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön

Deep Learning with Python, Second Edition, François Chollet

Applications of Deep Neural Networks with Keras, Jeff Heaton

Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David

Bibliografija

[1] Blackard, J. (1998). Covertype [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50K5N.

[2] Coops, N. C., Kearney, S. P., Bolton, D. K., & Radeloff, V. C. (2018). Remotely-sensed productivity clusters capture global biodiversity patterns. Scientific Reports, 8, 16261.

[3] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 226–231).

[4] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31 (3), 264–323.

[5] Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. Hydrology and Earth System Sciences, 24(3), 1081–1100.

[6] Legendre, P., & Legendre, L. (2012). Numerical Ecology (3rd ed.). Elsevier.

[7] Shetty, C., Seema, S., Sowmya, B. J., Nandalike, R, Supreeth, S., Dayananda, P., Rohith, S., Ranjan, R., & Goud, V. (2024). A Machine Learning Approach for Environmental Assessment on Air Quality and Mitigation Strategy. Journal of Engineering, 2024(1), 2893021.

[8] Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.

[9] Zou, K. (2022). The Dynamic Detection of Water Quality Monitoring and Pollution Prevention and Control. In Journal of Physics: Conference Series (Vol. 2152, No. 1, p. 012017). IOP Publishing.