Tema: Gručenje in analiza vzorcev v ekosistemih

Predmet: Strojno učenje za okoljske aplikacije

Učni izidi

Po zaključku te učne enote bodo študenti znali:

- Uporabiti tehnike gručenja za prepoznavanje vzorcev in skupin v okoljskih podatkovnih zbirkah.
- Ovrednotiti kakovost gruč z notranjimi in zunanjimi merami validacije.
- Razložiti ekološki pomen rezultatov nenadzorovanih modelov.
- Uporabiti rezultate gručenja za podporo ohranjanju narave in upravljanju ekosistemov.

Uvod

Gručenje je temeljni pristop nenadzorovanega učenja, ki podatkovne vzorce razvršča v skupine na podlagi podobnosti, ne da bi potreboval vnaprej določene oznake. Pri analizi okoljskih podatkov je še posebej dragoceno, saj ekosistemi pogosto izkazujejo kompleksne in večrazsežne strukture, v katerih jasne meje med razredi niso očitne. S prepoznavanjem naravnih skupin v podatkih gručenje podpira naloge, kot so ocenjevanje biotske raznovrstnosti, analiza kemijskih lastnosti tal in razlaga opazovanj z daljinskim zaznavanjem, pri tem pa upošteva tudi šum in nepravilne vzorce, ki so pri ekoloških meritvah pogosti [4,6,8]. Nenadzorovano gručenje podatkov o produktivnosti, pridobljenih z daljinskim zaznavanjem, je bilo povezano z vzorci biotske raznovrstnosti na velikih prostorskih območjih [2].

Gručenje z metodo k-sredin (K-Means)

Metoda k-sredin razdeli podatkovno zbirko na k gruč, tako da točke iterativno dodeljuje najbližjemu centroidu in posodablja položaje centroidov, da čim bolj zmanjša varianco znotraj gruč. Zaradi preprostosti in računske učinkovitosti je eden najpogosteje uporabljenih algoritmov gručenja.

Prednosti:

- Hitra metoda, ki jo je mogoče učinkovito uporabiti tudi pri velikih podatkovnih zbirkah.
- Preprosta za izvedbo in razlago.

Omejitve:

- Število gruč k je treba določiti vnaprej.
- Občutljiva na inicializacijo in osamelce.
- Najbolje deluje pri kroglastih, enako velikih gručah.

Uporaba na področju okolja: odkrivanje vzorcev onesnaženja v meritvah kakovosti vode [9].

Ključna omejitev algoritma k-sredin je zahteva, da se število gruč (*k*) določi vnaprej. Pri uporabi na področju okolja je dejansko število ekosistemskih režimov le redko znano vnaprej, neustrezna izbira *k* pa lahko privede do preveč poenostavljenih ali razdrobljenih predstavitev ekološke strukture.

Za reševanje te težave se pri izbiri *k* pogosto uporabljajo tehnike notranjega vrednotenja. Te metode ocenjujejo kakovost gručenja na podlagi strnjenosti in ločenosti gruč, ne da bi se opirale na zunanje oznake. Med pogosto uporabljenimi pristopi sta metoda komolca, ki preučuje zmanjševanje variance znotraj gruč ob povečevanju *k*, in analiza silhuete, ki ocenjuje, kako dobro se posamezna podatkovna točka ujema z dodeljeno gručo v primerjavi s sosednjimi gručami.

V praksi končna izbira *k* pogosto pomeni iskanje ravnotežja med kvantitativnimi rezultati validacije in možnostjo ekološke razlage, s čimer se zagotovi, da prepoznane gruče predstavljajo smiselne okoljske režime in ne zgolj matematičnih razdelitev.

Hierarhično gručenje

Hierarhično gručenje ustvari hierarhijo gruč, ki je pogosto prikazana z dendrogramom in razkriva gnezdene ravni podobnosti med vzorci. Lahko je združevalno (od spodaj navzgor) ali razdruževalno (od zgoraj navzdol). Ta metoda je še posebej uporabna pri ekoloških raziskavah, pri katerih je treba razumeti odnose na več prostorskih ravneh.

Prednosti:

- Števila gruč ni treba določiti vnaprej.
- Ustvari dendrograme, ki razkrivajo ekološke strukture na več prostorskih ravneh.
- Prilagodljivi kriteriji povezovanja (enojno, popolno, povprečno povezovanje).

Omejitve:

- Pri zelo velikih podatkovnih zbirkah je računsko zahtevno.
- Občutljivo na šum in skaliranje spremenljivk.

- Rezultate je lahko težko razložiti, če gruče niso dobro ločene.

Uporaba na področju okolja: gručenje hidroloških prispevnih območij na podlagi pretočnih in padavinskih režimov [5].

DBSCAN (prostorsko gručenje na podlagi gostote za podatke s šumom)

DBSCAN združuje točke, ki ležijo tesno skupaj, točke na območjih z nizko gostoto pa označi kot šum. V nasprotju z metodo k-sredin ne zahteva vnaprejšnje določitve števila gruč in lahko prepozna gruče poljubnih oblik.

Prednosti:

- Lahko odkrije gruče nepravilnih oblik.
- Samodejno prepozna šum/osamelce.
- Števila gruč ni treba določiti vnaprej.

Omejitve:

- Uspešnost je močno odvisna od izbire parametrov (epsilon, minPts).
- Ima težave pri gručah z različno gostoto.
- Pri zelo velikih podatkovnih zbirkah je manj skalabilen kot metoda k-sredin.

Uporaba na področju okolja: odkrivanje žarišč onesnaženja v omrežjih senzorjev za spremljanje kakovosti zraka ali vode [3, 7].

Mere za vrednotenje gruč

V nasprotju z nadzorovanim razvrščanjem gručenje pri večini okoljskih podatkovnih zbirk nima na voljo referenčnih oznak. Vrednotenje kakovosti gručenja zato temelji na indeksih notranjega ali zunanjega vrednotenja, ki količinsko opredeljujejo strnjenost, ločenost ali ujemanje s skupinami, ki jih določijo strokovnjaki.

Koeficient silhuete

- Meri, kako podobna je posamezna točka lastni gruči v primerjavi z drugimi gručami.
- Zavzema vrednosti od -1 do 1; višje vrednosti pomenijo dobro ločene in povezane gruče.
- Uporaben za izbiro števila gruč k pri metodi k-sredin.

Davies–Bouldinov indeks

- Razmerje med razpršenostjo znotraj gruč in ločenostjo med gručami.
- Nižje vrednosti pomenijo boljše gručenje.
- Računsko učinkovit in uporaben pri velikih okoljskih podatkovnih zbirkah.

Calinski–Harabaszev indeks

- Primerja varianco med gručami z varianco znotraj gruč.
- Višje vrednosti ustrezajo jasneje opredeljenim gručam.
- Pogosto se uporablja na področju ekologije pri podatkih z več spremenljivkami.

Prilagojeni Randov indeks (ARI)

- Mera zunanje validacije, ki rezultate gručenja primerja s podatki, ki so jih označili strokovnjaki, ali z znanimi ekološkimi regijami.
- Prilagojen je za naključno ujemanje, pri čemer 1 pomeni popolno ujemanje, 0 pa naključno razdelitev.
- Dragocen pri validaciji na podlagi uveljavljenih zemljevidov ekoregij ali naravovarstvenih enot.

Validacija gruč zagotavlja, da odkrite strukture niso artefakti algoritma, temveč odražajo smiselne ekološke vzorce. V okoljskih raziskavah kombiniranje notranjih indeksov (npr. koeficienta silhuete, Davies–Bouldinovega indeksa) z zunanjo validacijo na podlagi strokovno določenih razvrstitev zagotavlja trdne dokaze za ekološke razlage.

Študija primera: prepoznavanje ekosistemskih režimov z nenadzorovanim gručenjem okoljskih podatkov

Cilj

Ta študija primera prikazuje, kako lahko s tehnikami nenadzorovanega učenja v okoljskih podatkih odkrivamo naravne skupine in ekološke režime brez vnaprej določenih oznak. Cilj je prepoznati območja s podobnimi okoljskimi značilnostmi in te gruče razlagati kot različne ekosistemske ali okoljske režime. Takšno odkrivanje vzorcev je zlasti dragoceno pri ekološki analizi, kjer so naravne meje pogosto postopne, neostre ali neznane.

Uporabljene knjižnice

Študija primera je izvedena v Pythonu z uporabo široko uveljavljenih odprtokodnih knjižnic za analizo podatkov, strojno učenje in vizualizacijo:

- **NumPy**
  Uporablja se za numerične operacije, delo s polji in naključno vzorčenje.
- **Pandas**
  Uporablja se za nalaganje, obdelavo in predobdelavo tabelaričnih zbirk okoljskih podatkov.
- **Scikit-learn (sklearn).**
  Uporablja se za:
  - predobdelavo podatkov (StandardScaler),
  - algoritme gručenja (metoda k-sredin),
  - zmanjševanje razsežnosti (PCA),
  - notranje mere vrednotenja gruč (koeficient silhuete, Davies–Bouldinov indeks).
- **SciPy**
  Uporablja se za hierarhično gručenje in prikaz dendrogramov.
- **Matplotlib**
  Uporablja se za prikaz rezultatov gručenja, razsevnih diagramov, projekcij PCA in dendrogramov.

Ta izbira podpira preglednost in omogoča preprosto izvedbo postopka v standardnih okoljih, kot sta Jupyter Notebook in Google Colab.

Vir podatkov

Analiza temelji na zbirki podatkov Forest Cover Type (Covertype) [1], ki je prosto dostopna v repozitoriju UCI Machine Learning Repository. Zbirka vsebuje okoljske in topografske meritve, zbrane na gozdnatih območjih nacionalnega gozda Roosevelt v ZDA. Čeprav zbirka vključuje oznako pokrovnosti tal, so te informacije v tej študiji primera namerno izključene, da se ohrani nenadzorovana narava analize.

Opis podatkov

Vsako opazovanje predstavlja geografsko lokacijo, opisano z okoljskimi spremenljivkami, kot so:

- Nadmorska višina
- Naklon
- Ekspozicija
- Vodoravna in navpična oddaljenost od vodovja
- Oddaljenost od cest
- Indeksi senčenja reliefa

Te spremenljivke skupaj opisujejo relief, gradiente vlažnosti in dostopnost, ki so ključni dejavniki razlikovanja med ekosistemi.

Predobdelava podatkov

Za pripravo zbirke podatkov na gručenje se izvedejo naslednji koraki:

- Izbor samo zveznih okoljskih in topografskih spremenljivk
- Odstranitev oznake tipa gozdnega pokrova
- Standardizacija spremenljivk na povprečje nič in varianco ena, da se zagotovi njihov enak prispevek pri metodah gručenja, ki temeljijo na razdalji

```python
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()
```

Prevod oznak in komentarjev: »Naloži podatkovno zbirko (različica Covertype v obliki CSV)«; »Odstrani oznako tipa gozdnega pokrova (zadnji stolpec)«; »Standardiziraj značilke«.

Metode gručenja

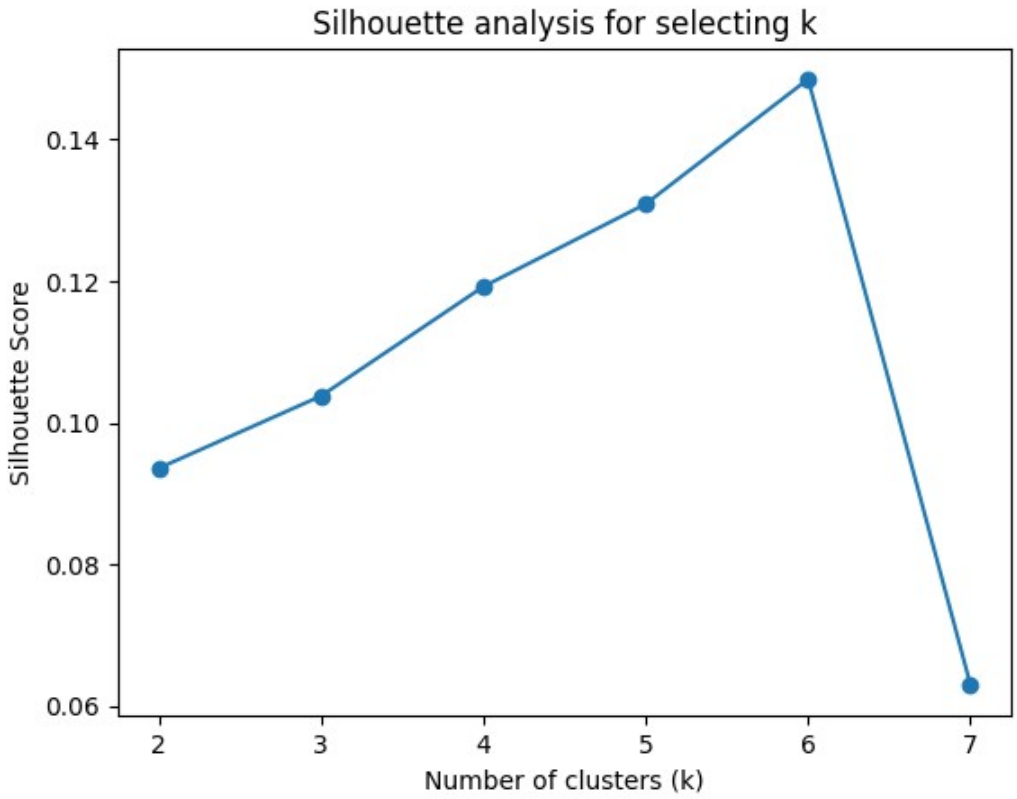
Uporabita se dve dopolnjujoči se tehniki gručenja:

Gručenje z metodo k-sredin

Metoda k-sredin se uporablja za razdelitev zbirke podatkov na vnaprej določeno število gruč glede na okoljsko podobnost. Pred uporabo metode k-sredin je bilo število gruč preučeno s tehnikami notranjega vrednotenja. Analiza silhuete je bila uporabljena za oceno kakovosti gručenja pri različnih vrednostih *k*. Te analize podpirajo izbiro števila gruč, ki uravnoteži ločenost, strnjenost in možnost ekološke razlage.

```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()
```

Prevod oznak in komentarjev: komentarja »Razpon vrednosti k za preizkus« in »Izriši rezultate«; izpis »Izračun koeficienta silhuete za k = {k} ...«; naslov grafa »Analiza silhuete za izbiro k (celotna podatkovna zbirka)«; os x »Število gruč (k)«; os y »Koeficient silhuete«.



Prevod oznak in komentarjev: naslov grafa »Analiza silhuete za izbiro k«; os x »Število gruč (k)«; os y »Koeficient silhuete«.

Analiza silhuete je pokazala višje vrednosti pri k = 6, kar s povsem statističnega vidika kaže na boljšo ločenost gruč. Vendar pri okoljskih uporabah optimalnega števila gruč ne določajo zgolj notranje mere vrednotenja. Ekosisteme oblikujejo zvezni gradienti, ne pa ostro ločeni razredi, povečanje števila gruč pa lahko ekološko povezane režime razdrobi na manjše enote, ki jih je težje razlagati.

Zato je bilo izbrano manjše število gruč (k = 4), da bi uravnotežili statistično kakovost gručenja z možnostjo ekološke razlage in pedagoško jasnostjo. Ta izbira omogoča prepoznavanje širših in smiselnih ekosistemskih režimov, ki se ujemajo s prevladujočimi okoljskimi gradienti, hkrati pa preprečuje pretirano razdrobljenost, ki bi lahko otežila praktično razumevanje in uporabo

```
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()
```

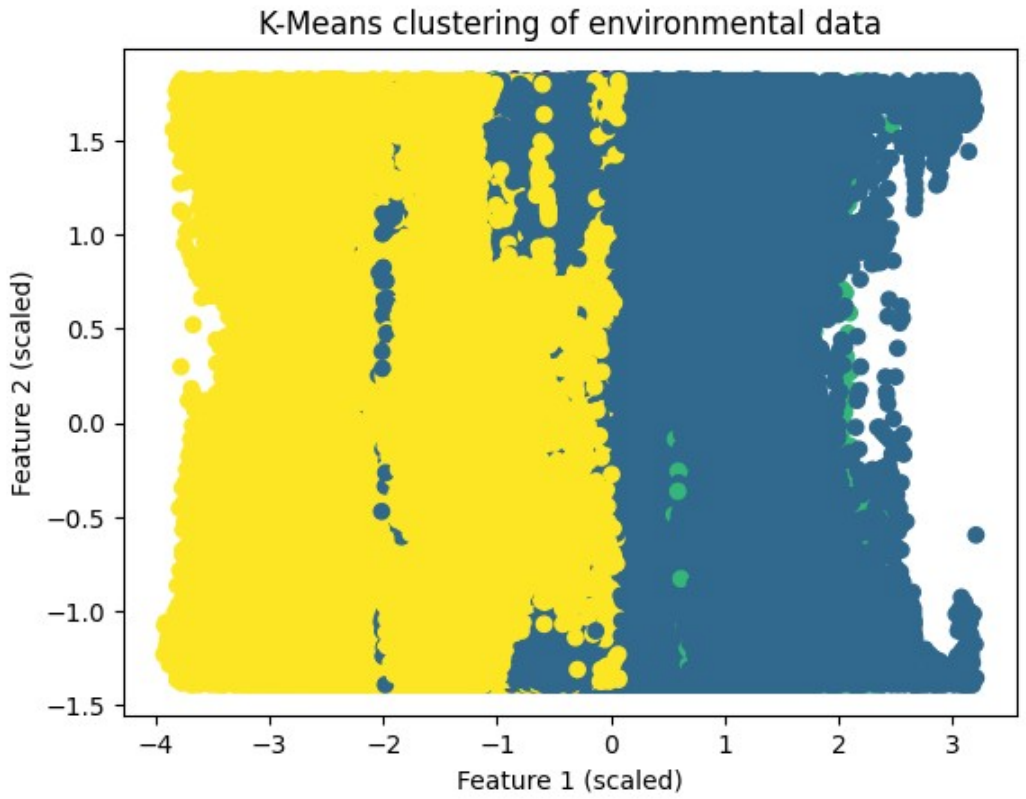
✓ 1.9s

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

Prevod oznak in komentarjev: komentarja »Uporabi metodo k-sredin« in »Dodaj oznake gruč v podatkovni okvir«; glava izpisa »Gruča« (Cluster); zaključna vrstica izpisa »Ime: count, podatkovni tip: int64«.

```
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()
```

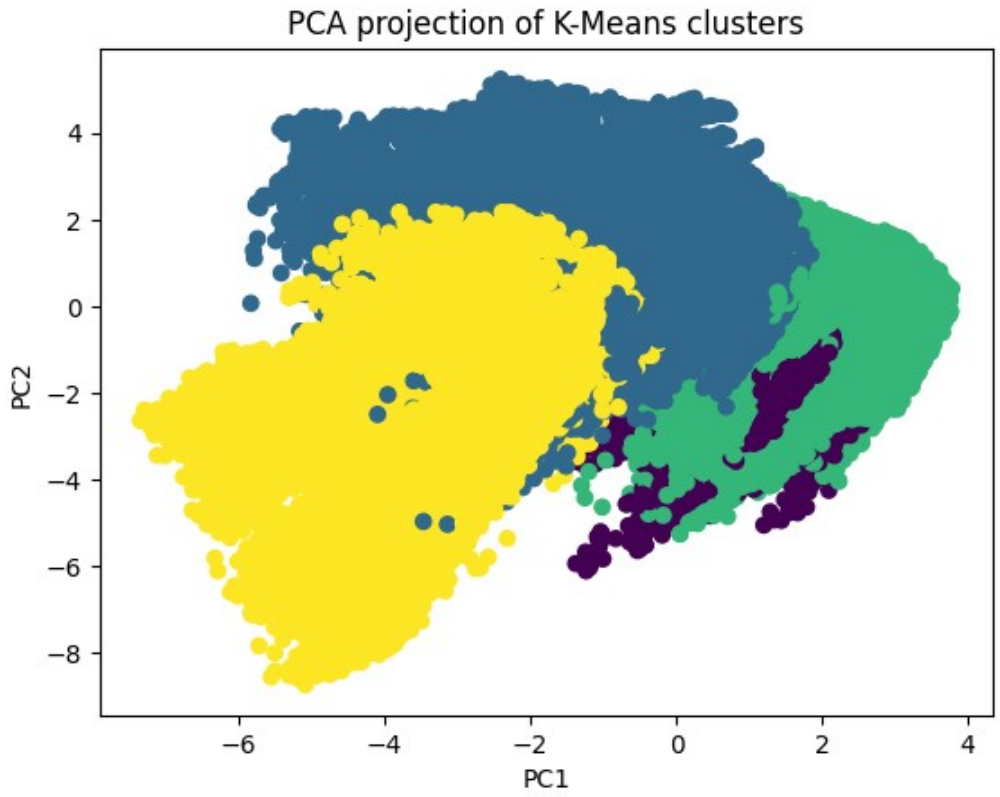
Prevod oznak in komentarjev: naslov grafa »Gručenje okoljskih podatkov z metodo k-sredin«; os x »Značilka 1 (skalirana)«; os y »Značilka 2 (skalirana)«.



Prevod oznak in komentarjev: naslov grafa »Gručenje okoljskih podatkov z metodo k-sredin«; os x »Značilka 1 (skalirana)«; os y »Značilka 2 (skalirana)«.

```
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()
```

Prevod oznak in komentarjev: komentar »Uporabi PCA za zmanjšanje na 2 razsežnosti«; naslov grafa »Projekcija gruč metode k-sredin s PCA«.



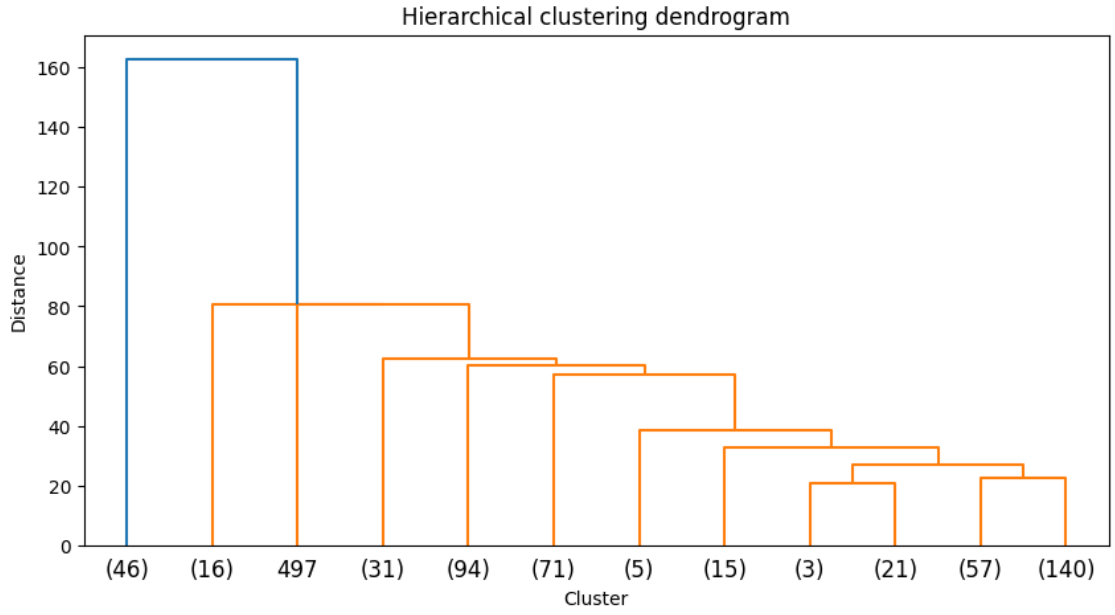
Prevod oznak in komentarjev: naslov grafa »Projekcija gruč metode k-sredin s PCA«.

Hierarhično gručenje

Hierarhično aglomerativno gručenje se uporablja za raziskovanje gnezdenih razmerij med vzorčnimi lokacijami. Dendrogrami se uporabljajo za prikaz ekoloških struktur v različnih merilih in razkrivajo, kako se lokalni vzorci združujejo v širše ekosistemske skupine.

```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()
```

Prevod oznak in komentarjev: komentar »Za vizualizacijo uporabi podmnožico«; naslov grafa »Dendrogram hierarhičnega gručenja«; os x »Gruča«; os y »Razdalja«.



Prevod oznak in komentarjev: naslov grafa »Dendrogram hierarhičnega gručenja«; os x »Gruča«; os y »Razdalja«.

Vrednotenje gruč

Kakovost gruč se ocenjuje z indeksi notranjega vrednotenja:

- Koeficient silhuete
- Davies–Bouldinov indeks

Te mere pomagajo zagotoviti, da so prepoznane gruče strnjene, dobro ločene in niso artefakti naključnega razdeljevanja.

```
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 notranjih mer vrednotenja, kot sta koeficient silhuete in Davies–Bouldinov indeks, je lahko pri obsežnih zbirkah okoljskih podatkov računsko zahteven, saj te mere temeljijo na izračunih razdalj med pari. Glede na velikost zbirke podatkov, uporabljene v tej študiji primera, je bilo vrednotenje kakovosti gručenja zato izvedeno na reprezentativni naključni podmnožici podatkov. Ta pristop ohranja osnovno strukturo podatkov, hkrati pa zagotavlja računsko izvedljivost in je običajna praksa pri analizah gručenja obsežnih podatkov. Dobljene mere omogočajo zanesljivo primerjavo kakovosti gručenja, ne da bi bil potreben izčrpen izračun na celotni zbirki podatkov.

```
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
```

Prevod oznak in komentarjev: »Za hitrejše vrednotenje uporabi podvzorec«.

Izračun notranjih mer vrednotenja, kot sta koeficient silhuete in Davies–Bouldinov indeks, je lahko pri obsežnih zbirkah okoljskih podatkov računsko zahteven, saj te mere temeljijo na izračunih razdalj med pari. Glede na velikost zbirke podatkov, uporabljene v tej študiji primera, je bilo vrednotenje kakovosti gručenja zato izvedeno na reprezentativni naključni podmnožici podatkov. Ta pristop ohranja osnovno strukturo podatkov, hkrati pa zagotavlja računsko

izvedljivost in je običajna praksa pri analizah gručenja obsežnih podatkov. Dobljene mere omogočajo zanesljivo primerjavo kakovosti gručenja, ne da bi bil potreben izčrpen izračun na celotni zbirki podatkov.

```
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
```

Prevod oznak in komentarjev: »Za hitrejše vrednotenje uporabi podvzorec«.

Razlaga in ekološki pomen

Dobljene gruče je mogoče razlagati kot različne ekosistemske režime, na primer:

- Gozdna okolja na nizkih v primerjavi z visokimi nadmorskimi višinami
- Vlažne v primerjavi s suhimi razmerami na terenu
- Dostopna v primerjavi z odročnimi gorskimi območji

S primerjavo gruč z znanimi okoljskimi gradienti lahko študenti vsako skupino povežejo z značilnimi ekološkimi razmerami namesto z vnaprej določenimi razredi pokrovnosti tal.

Pomen za upravljanje ekosistemov

Odkrite gruče zagotavljajo objektivno podlago za:

- Razvrščanje in kartiranje ekosistemov
- Prepoznavanje okoljsko podobnih območij za načrtovanje ohranjanja narave
- Podporo odločitvam o upravljanju rabe zemljišč na območjih s pomanjkljivimi podatki ali slabo kartiranih območjih

Ta študija primera prikazuje, kako gručenje dopolnjuje strokovno znanje, saj na podlagi podatkov razkriva ekološke strukture, ki se morda ne ujemajo z upravnimi mejami.

Naloge za vajo

- Prenesite podatkovno zbirko o tipih gozdnega pokrova (Covertype) iz repozitorija UCI Machine Learning Repository [1].
- Raziščite podatkovno zbirko in izvedite začetno opisno analizo okoljskih in topografskih spremenljivk.
- Izberite podmnožico ustreznih zveznih spremenljivk (npr. nadmorsko višino, naklon, oddaljenost od vodnih virov) in utemeljite izbor.
- Standardizirajte izbrane spremenljivke in preučite, kako spreminjanje merila vpliva na gručenje, ki temelji na razdaljah.
- S tehnikami notranjega vrednotenja (metodo komolca in analizo silhuete) raziščite različne vrednosti števila gruč (k).
- Izvedite gručenje z metodo k-sredin z izbrano vrednostjo k in analizirajte dobljeno strukturo gruč.
- Izvedite hierarhično gručenje na manjši podmnožici podatkov in z dendrogramom prikažite odnose na različnih ravneh podrobnosti.
- Primerjajte rezultate gručenja z metodo k-sredin in hierarhičnimi metodami ter poudarite podobnosti in razlike.
- Gruče interpretirajte z vidika prevladujočih okoljskih gradientov in ne vnaprej določenih razredov pokrovnosti tal.
- Razmislite o omejitvah gručenja pri analizi ekosistemov in razpravljajte o tem, kako bi rezultati lahko dopolnili strokovno klasifikacijo ekosistemov in odločitve o upravljanju.

Preverjanje znanja

Kaj je glavni namen uporabe tehnik gručenja pri analizi ekosistemskih vzorcev?

A. Napovedovanje prihodnjih vrednosti okoljskih spremenljivk
B. Razvrščanje vzorcev v vnaprej določene razrede tveganja
C. Prepoznavanje naravnih skupin in vzorcev v neoznačenih okoljskih podatkih
D. Zmanjševanje razsežnosti zgolj za vizualizacijo

Katera omejitev je najbolj značilna za algoritem gručenja z metodo k-sredin pri okoljskih podatkovnih zbirkah?

A. Nezmožnost obdelave velikih podatkovnih zbirk
B. Zahteva po vnaprejšnji določitvi števila gruč (k)
C. Nezmožnost dela s številskimi spremenljivkami
D. Nezmožnost prepoznavanja strnjenih gruč

Katera metoda gručenja je najprimernejša za zaznavanje žarišč onesnaženja nepravilnih oblik ob hkratnem prepoznavanju šuma?

A. Metoda k-sredin (K-Means)
B. Hierarhično gručenje
C. DBSCAN
D. Logistična regresija

Zakaj pri analizi gručenja pogosto uporabljamo mere notranjega vrednotenja, kot je koeficient silhuete?

A. Ker so pri gručenju vedno na voljo referenčne oznake
B. Za napovedovanje prihodnjih sprememb ekosistemov
C. Za vrednotenje strnjenosti in ločenosti gruč brez označenih podatkov
D. Za optimizacijo nadzorovanih klasifikatorjev

Kako lahko rezultati gručenja podprejo odločitve o upravljanju in ohranjanju ekosistemov?

A. Z nadomeščanjem ekoloških terenskih opazovanj
B. Z zagotavljanjem natančnih napovedi sprememb ekosistemov
C. Z razkrivanjem funkcionalnih skupin in ekoloških gradientov, pomembnih za upravljanje
D. Z odpravljanjem negotovosti v okoljskih podatkih

Dodatni viri

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

Deep Learning with Python, 2. izdaja, François Chollet

Applications of Deep Neural Networks with Keras, Jeff Heaton

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

Literatura

[1] Blackard, J. (1998). Covertype [podatkovna zbirka]. 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. V: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD) (str. 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 (3. izdaja). 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 (2. izdaja). Pearson.

[9] Zou, K. (2022). The Dynamic Detection of Water Quality Monitoring and Pollution Prevention and Control. V: Journal of Physics: Conference Series (let. 2152, št. 1, str. 012017). IOP Publishing.