# Gručenje

Gručenje podatkov je metoda nenadzorovanega učenja, ki omogoča iskanje skritih vzorcev v podatkih brez vnaprej določenih oznak ali razredov. Podatke razvršča v gruče, kjer so si primeri znotraj posamezne skupine čim bolj podobni, medtem ko so različne gruče med seboj čim bolj različne. To omogoča boljše razumevanje strukture podatkov, njihovo organizacijo ter zmanjšanje dimenzionalnosti, kar olajša nadaljnjo analizo.

Postopki gručenja se pogosto uporabljajo pri analizi uporabnikov, kjer podjetja prepoznavajo različne segmente kupcev na podlagi njihovih nakupnih navad. V biomedicini omogočajo odkrivanje podtipov bolezni ali razvrščanje genetskih podatkov, v računalniškem vidu pa se uporabljajo za segmentacijo slik in prepoznavanje objektov. Prav tako so ključni pri naravnem jeziku, saj omogočajo avtomatsko razvrščanje dokumentov po temah, kar se pogosto uporablja pri iskalnikih in analizi besedil.

Poleg odkrivanja skritih struktur gručenje pomaga tudi pri odkrivanju anomalij, saj lahko izolirane točke v podatkih nakazujejo goljufive transakcije, napake v meritvah ali redke dogodke. Uporablja se v geografskih analizah za identifikacijo gostotnih območij, pri analizi omrežij za iskanje povezanih skupnosti ter v sistemih priporočanja, kjer omogoča prilagojeno ponudbo vsebin uporabnikom. Kot eno temeljnih orodij za analizo neoznačenih podatkov je gručenje nepogrešljivo v številnih disciplinah.

## Poskus formalnega zapisa cilja gručenja

Gručenje je metoda nenadzorovanega učenja, katere cilj je poiskati optimalno razdelitev podatkov v k gruč, tako da so podatki znotraj posamezne gruče čim bolj podobni, podatki med različnimi gručami pa čim bolj različni. Formalno lahko gručenje opredelimo kot postopek, ki množico podatkovnih primerov $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ razdeli v $k$ disjunktnih gruč $C_1, C_2, ..., C_k$, pri čemer velja:

1. Vsak podatkovni primer pripada natanko eni gruči:

$$C_1 \cup C_2 \cup ... \cup C_k = X, \quad C_i \cap C_j = \emptyset, \quad \text{za } i \neq j.$$

2. Znotrajgručna podobnost je čim večja:

$$\arg\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{\mathbf{x}\in C_i} d(\mathbf{x}, \mu_i),$$

kjer je $\mu_i$ središče gruče $C_i$ in $d(\mathbf{x}, \mu_i)$ funkcija razdalje med primerom $\mathbf{x}$ in središčem gruče (npr. Evklidska razdalja).

3. Medgručna različnost je čim večja:

$$\arg\max_{C_1,\ldots,C_k} \sum_{i\neq j} d(\mu_i, \mu_j),$$

kjer $d(\mu_i, \mu_j)$ meri razdaljo med središči različnih gruč, da zagotovimo njihovo dobro ločljivost.

Praktične metode gručenja optimizirajo različne cilje. Tehnika voditeljev, na primer, minimizira vsoto kvadratnih znotrajgručnih razdalj, hierarhično gručenje gradi hierarhijo gruč glede na razdaljo med skupinami, DBSCAN pa temelji na iskanju gostotnih območij brez potrebe po določitvi števila gruč vnaprej.

# Vrste gručenja

Metode gručenja podatkov se razlikujejo glede na način oblikovanja gruč in kriterije, ki jih uporabljajo za združevanje podatkovnih primerov. Različni algoritmi so prilagojeni različnim tipom podatkov in ciljem analize. Na splošno jih lahko razdelimo v šest glavnih skupin:

## Particijsko gručenje

Particijske metode razdelijo podatkovno množico na fiksno število $k$ gruč, pri čemer vsaka podatkovna točka pripada natanko eni gruči. Cilj je minimizirati znotrajgručne razdalje in optimizirati razporeditev podatkovnih primerov v gruče. Najbolj znan predstavnik je K-means, ki temelji na iterativnem iskanju centrov gruč in ponovnem razporejanju točk. Podoben je K-medoids, ki namesto povprečja uporablja dejanske podatkovne primere kot središča gruč, s čimer je manj občutljiv na osamelce. Za primere, kjer lahko ena podatkovna točka pripada več gručam hkrati, pa se uporablja Fuzzy C-means, ki vsaki točki dodeli verjetnostno pripadnost več gručam. Particijske metode so računsko učinkovite in enostavne za implementacijo,

vendar zahtevajo vnaprejšnjo izbiro števila gruč in so občutljive na izbiro začetnih centrov.

## Hierarhično gručenje

Hierarhične metode gradijo dendrogram, drevesno strukturo gruč, ki omogoča analizo podatkov na različnih nivojih podrobnosti. Postopek je lahko agregativen (spodaj navzgor), kjer se posamezne točke postopoma združujejo v večje gruče, ali diviziven (od zgoraj navzdol), kjer se celotna množica postopoma deli na manjše gruče. Različne metode merjenja razdalj med gručami, kot so single-linkage (najmanjša razdalja med primeri), complete-linkage (največja razdalja med primeri) in Wardova metoda (minimizacija variance), vplivajo na obliko gruč. Prednost hierarhičnega gručenja je, da ne zahteva predhodne določitve števila gruč in omogoča vizualno interpretacijo rezultatov, vendar je računsko zahtevno in manj primerno za velike podatkovne množice.

## Gručenje na podlagi gostote

Gostotne metode gručenja ne zahtevajo vnaprej določenega števila gruč, temveč gruče identificirajo kot gosto poseljena območja v podatkovnem prostoru, ločena z območji nizke gostote. Najbolj znan predstavnik je DBSCAN (Density-Based Spatial Clustering of Applications with Noise), ki določa gruče glede na število točk v določenem radiju ter omogoča identifikacijo osamelcev. Njegova razširitev OPTICS omogoča zaznavanje gruč različnih gostot. Mean-Shift je še en pristop, ki iterativno premika središča gruč proti gostim območjem. Ti algoritmi so uporabni pri iskanju nepravilnih oblik gruč in odkrivanju anomalij, vendar imajo težave pri podatkih z neenakomerno gostoto in so računsko zahtevnejši.

## Modelno temelječe gručenje

Metode, ki temeljijo na statističnih modelih, predpostavljajo, da so podatki generirani iz kombinacije več skritih porazdelitev, običajno normalnih porazdelitev. Gaussian Mixture Models (GMM) modelira gruče kot kombinacijo Gaussovih porazdelitev in uporablja pričakovalno-maksimizacijski (EM) algoritem za iskanje optimalnih parametrov. Naprednejši pristopi, kot so Bayesovi pristopi, omogočajo določanje števila gruč na podlagi posteriornih verjetnosti. Modelno temelječe metode omogočajo mehko gručenje, kjer lahko posamezen primer pripada več gručam, in so še posebej uporabne, kadar imajo gruče različne oblike in velikosti. Njihova slabost je potreba po zapleteni nastavitvi parametrov in računska zahtevnost.

# Gručenje na grafih

Grafovski pristopi obravnavajo podatke kot graf, kjer so podatkovni primeri vozlišča, povezave med njimi pa temeljijo na podobnosti. Spectral Clustering uporablja lastne vrednosti Laplaceove matrike za iskanje gruč in je primeren za kompleksne podatkovne strukture, ki niso nujno konveksne. Louvain Method se pogosto uporablja za iskanje skupnosti v velikih omrežjih, kjer optimizira modularnost povezav. Grafovske metode so še posebej uporabne pri analizi socialnih omrežij, molekularnih struktur in geometrijsko kompleksnih podatkovnih prostorov, vendar so pri velikih podatkovnih množicah lahko računsko zahtevne.

| Vrsta gručenja | Opis | Primeri algoritmov |
|---|---|---|
| Particijsko | Razdeli podatke na fiksno $k$ število gruč | K-means, K-medoids, Fuzzy C-means |
| Hierarhično | Gradi drevesno strukturo gruč | Single-linkage, Ward, Complete-linkage |
| Gostotno | Gruče so območja visoke gostote podatkov | DBSCAN, OPTICS, Mean-Shift |
| Modelno | Gruče temeljijo na statističnih modelih | GMM, Bayesovi pristopi |
| Na grafih | Gruče temeljijo na teoriji grafov | Spectral Clustering, Louvain |

Različni pristopi k gručenju ponujajo različne prednosti in slabosti, odvisno od značilnosti podatkov in želenega izida analize. Particijske metode so učinkovite in enostavne, a zahtevajo vnaprej določeno število gruč. Hierarhične metode omogočajo večnivojsko analizo, vendar so računsko zahtevne. Gostotno gručenje je primerno za podatke z nelinearnimi strukturami, modelno temelječe metode omogočajo fleksibilno dodelitev primerov več gručam, grafovsko gručenje pa je uporabno za kompleksne odnose med podatki. Mehko gručenje je posebej uporabno tam, kjer so meje med skupinami nejasne. Izbira optimalne metode je odvisna od narave podatkov in analitičnih ciljev.

# Nekaj izbranih pristopov

Med zgornjimi pristopi tu malce podrobneje poglejmo nekaj izbranih in morda najbolj pogosto uporabljanih.

## Hierarhično gručenje

Hierarhično gručenje razvrsti podatke v drevesno strukturo gruč (*dendrogram*), kjer je vsaka podatkovna točka sprva obravnavana kot samostojna gruča, nato pa se postopoma združujejo v večje skupine, dokler ne ostane ena sama gruča ali pa se doseže določena stopnja združevanja. Obstajata dva pristopa: združevalno gručenje (bottom-up), kjer se začne z posameznimi točkami in jih postopoma združujemo v večje gruče, ter divizivno gručenje (top-down), kjer se začne z eno samo gručo in jo postopoma delimo na manjše podskupine. Najbolj uporabljano je združevalno gručenje, morda tudi zaradi preprostosti algoritma (glej spodaj). Prednost hierarhičnega gručenja je, da omogoča analizo podatkov na različnih granularnih nivojih, slabost pa je visoka računska ali pa spominska zahtevnost ($O(n^2)$ ali več) in občutljivost na šum.

Združevalno gručenje lahko izvedemo s spodnjim algoritmom:

1. Vsako podatkovno točko obravnavamo kot samostojno gručo.
2. Izračunamo matriko razdalj med vsemi pari gruč.
3. Združimo najbolj podobni gruči v novo, večjo gručo.
4. Posodobimo matriko razdalj (upoštevajoč izbrani način ocenjevanja podobnosti).
5. Ponovimo korake 3 in 4, dokler ne ostane ena sama gruča ali dokler ne dosežemo želenega števila gruč.

Hierarhično gručenje uporablja različne metode za merjenje razdalje med gručami. Izbira pravilne metode povezovanja (*linkage*) vpliva na strukturo in interpretacijo rezultatov gručenja. Naj bosta $C_i$ in $C_j$ gruči, $d(\mathbf{x}_a, \mathbf{x}_b)$ pa razdalja med točkama $\mathbf{x}_a \in C_i$ in $\mathbf{x}_b \in C_j$. Matematično so načini ocenjevanja podobnosti gruč definirani kot:

1. Najbližji sosed (angl. single-linkage). Razdalja med gručama je določena kot najmanjša razdalja med katerima koli dvema točkama iz različnih gruč. Omogoča identifikacijo nepravilnih oblik gruč, a je občutljivo na verižni učinek (*chaining*).

$$d(C_i, C_j) = \min_{\mathbf{x}_a \in C_i, \mathbf{x}_b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b)$$

2. Najbolj oddaljen sosed (complete-linkage). Razdalja med gručama je določena kot največja razdalja med katerima koli dvema točkama. Gradi bolj kompaktne gruče, a je manj odporna na osamelce.

$$d(C_i, C_j) = \max_{\mathbf{x}_a \in C_i, \mathbf{x}_b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b)$$

    - Združuje gruče na podlagi največje razdalje med katerima koli dvema točkama.
    - Ustvari bolj kompaktne in kroglaste gruče, vendar je občutljivo na osamelce.
3. Povprečna razdalja (average-linkage). Povprečna razdalja med vsemi pari točk v različnih gručah. Omogoča uravnoteženo povezovanje gruč in je manj občutljiva na osamelce.

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b)$$

    - Združuje gruče na podlagi povprečne razdalje med vsemi pari točk.
    - Uravnotežena metoda, ki daje bolj stabilne rezultate.
4. Minimalno povečanje variance (Wardova metoda).. Minimizira povečanje variance pri združevanju gruč, kar vodi do bolj enakomerno velikih gruč. Pogosto se uporablja, kadar so podatki razporejeni v približno sferične gruče.

$$d(C_i, C_j) = \sum_{\mathbf{x} \in C_{ij}} \|\mathbf{x} - \mu_{ij}\|^2 - \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2 - \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|^2$$

    - Združuje gruče tako, da minimizira povečanje znotrajgručne variance.
    - Primerna za podatke, kjer so gruče približno sferične.

Vsaka metoda ima svoje prednosti in slabosti, izbira pa je odvisna od strukture podatkov in aplikacije. Tipično, sicer, pri nekih podatkih iz realnega, npr. poslovnega sveta, bomo uporabili ali Evklidsko ali kosinusno razdaljo v kombinaciji z Wardovo metodo.

Glavna prednost hierarhičnega združevanja v skupine je grafična predstavitev rezultatov z dendrogramom, ki ga je mono lepo kombinirati z ostalimi predstavitvami podatkov, npr. s toplotno karto. Pomanlkljivosti pa je več: dendrogrami so primerni samo za predstavitev manjših podatkov, npr. do največ nekaj sto primerov, njhova predstavitev ni enolična (za isto gručenje je možno sestaviti eksponentno mnogo dendrogramov), odločitev, koliko skupin je v

podatkih, pa je prepuščena uporabniku.

## Metoda voditeljev

Gručenje po metodi voditeljev (K-means) je ena najpogosteje uporabljenih metod particijskega gručenja. Algoritem deli podatkovno množico na $k$ gruč tako, da minimizira vsoto kvadratnih razdalj med podatkovnimi primeri in pripadajočimi središči gruč (*centroidi*). Vsaka podatkovna točka je dodeljena gručam glede na najbližje središče, nato pa se središča posodobijo kot povprečne vrednosti vseh točk v posamezni gruči. Postopek se ponavlja, dokler središča ne konvergirajo ali se spremembe v dodelitvah ustavijo. Zaradi svoje preprostosti in učinkovitosti je metoda široko uporabljena v različnih domenah, kot so analiza uporabnikov, razvrščanje slik in obdelava biomedicinskih podatkov.

Matematično je cilj metode minimizirati znotrajgručno napako, podano kot:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2$$

kjer je $C_i$ množica podatkovnih primerov v $i$-ti gruči, $\mu_i$ pa središče gruče, izračunano kot povprečje vseh točk v njej. Algoritem se izvaja v dveh ključnih korakih: (1) Dodelitev točk – vsaka točka je dodeljena najbližjemu središču gruče na podlagi Evklidske razdalje, (2) Posodobitev središč – središča se premaknejo v povprečne vrednosti točk v gruči. Algoritem ponavlja ta dva koraka, dokler se središča ne stabilizirajo.

Čeprav je K-means učinkovit in hitro konvergira, ima nekatere pomanjkljivosti. Potrebno je vnaprej določiti število gruč $k$, algoritem pa je občutljiv na začetno izbiro središč, kar lahko vodi v lokalne minimume. Poleg tega slabo deluje pri podatkih z nepravilnimi oblikami gruč in je občutljiv na osamelce. Za izboljšanje stabilnosti se pogosto uporablja večkratno inicializacijo (K-means++), ki pametneje izbere začetna središča, ali alternativni algoritmi, kot so K-medoids, ki uporabljajo dejanske podatkovne točke kot središča gruč. Metoda je tudi občutljiva na osamelce, saj ti močno vplivajo na središča gruč, zato je pri podatkih s šumom pogosto bolje uporabiti gručenje z medoidi ali gostotne metode, kot je DBSCAN.

K-means++ je izboljšana inicializacija za K-means, ki zmanjšuje tveganje slabe konvergence v lokalne minimume zaradi slabo izbranih začetnih središč. Klasični K-means naključno izbere začetna središča, kar lahko povzroči neuravnotežene gruče ali počasno konvergenco. K-

means++ to izboljša tako, da inicializacijo središč izvede na podlagi razpršenosti podatkov, kar omogoča hitrejše in stabilnejše gručenje. Glavna prednost metode je, da zagotavlja boljšo porazdelitev začetnih centrov, kar poveča natančnost in zmanjšuje število iteracij, potrebnih za konvergenco.

Postopek inicializacije K-means++:

1. Naključno izberemo prvo središče $\mu_1$ izmed podatkovnih točk.
2. Za vsako točko $\mathbf{x}$ izračunamo njeno razdaljo do najbližjega že izbranega središča $D(\mathbf{x}) = \min_i d(\mathbf{x}, \mu_i)$.
3. Novo središče $\mu_j$ izberemo naključno, pri čemer je verjetnost izbire točke $\mathbf{x}$ sorazmerna z $D(\mathbf{x})^2$ (točke, ki so bolj oddaljene, imajo večjo možnost izbire).
4. Postopek ponavljamo, dokler ne izberemo vseh $k$ središč.
5. Ko so začetna središča določena, nadaljujemo s klasičnim K-means algoritmom (dodeljevanje točk in posodabljanje središč).

Ta metoda zagotavlja, da so začetna središča bolje razporejena po podatkovnem prostoru, kar povečuje natančnost in zmanjšuje občutljivost algoritma na začetno izbiro. Eksperimentalno se je pokazalo, da K-means++ v povprečju zahteva manj iteracij in zagotavlja rešitve, ki so bližje globalnemu optimumu v primerjavi s klasičnim K-means algoritmom.

## Medoidna metoda

Medoidna metoda gručenja (K-medoids) je različica metode z voditelji, kjer so središča gruč (*medoidi*) dejanske podatkovne točke namesto povprečij. Algoritem iterativno posodablja medoide tako, da minimizira vsoto razdalj med podatkovnimi primeri in pripadajočimi medoidi, kar ga naredi manj občutljivega na osamelce v primerjavi s K-means. Poleg tega omogoča uporabo poljubnih metričnih razdalj, ne le Evklidske, in je primernejši za ne-sferične gruče, saj ne predpostavlja enake velikosti in oblike gruč. Njegova glavna slabost je višja računska zahtevnost ($O(n^2)$) v primerjavi s K-means ($O(nk)$), kar ga omejuje pri zelo velikih podatkovnih množicah.

## Metoda DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je metoda gručenja, ki temelji na gostoti podatkov in ne zahteva vnaprej določenega števila gruč. Namesto tega

identificira gruče kot območja z visoko gostoto podatkovnih primerov, ločena z območji nizke gostote. To omogoča zaznavanje gruč nepravilnih oblik, kar je prednost pred metodami, kot je K-means, ki predpostavljajo sferične gruče. Poleg tega lahko DBSCAN identificira osamelce (točke, ki ne pripadajo nobeni gruči), zaradi česar je uporaben pri analizi anomalij, odkrivanju vzorcev v geolokacijskih podatkih in obdelavi velikih nestrukturiranih podatkovnih množic.

Algoritem DBSCAN temelji na kriteriju gostote, ki določa, da mora vsaka gruča vsebovati vsaj $\mathrm{minPts}$ točk v okolici radija $\varepsilon$ (*epsilon neighborhood*). Algoritem najprej izbere naključno podatkovno točko in preveri, koliko točk je znotraj njenega radija $\varepsilon$. Če jih je vsaj $\mathrm{minPts}$, je točka označena kot jedrna točka in začne se širjenje gruče – vse točke znotraj njenega dosega postanejo del iste gruče, nato pa se postopek rekurzivno ponovi za nove jedrne točke. Točke, ki imajo manj kot $\mathrm{minPts}$ sosedov, so lahko mejne točke (pripadajo gruči, vendar niso jedrne) ali osamelci (ne pripadajo nobeni gruči). Algoritem se nadaljuje, dokler niso vse točke dodeljene gručam ali označene kot osamelci.

DBSCAN ima več prednosti: ne zahteva vnaprej določenega števila gruč, zazna gruče nepravilnih oblik, identificira osamelce in dobro deluje pri velikih podatkovnih množicah. Slabosti metode so njena občutljivost na parametra $\varepsilon$ in minPts, ki ju je treba ustrezno nastaviti za različne podatkovne množice. Prav tako se DBSCAN slabo obnese pri podatkih z neenakomerno gostoto, saj lahko preveč razpršene gruče ostanejo nepovezane ali pa se več ločenih gostih območij združi v eno gručo.

Optimalne vrednosti parametrov $\varepsilon$ (epsilon) in $\mathrm{minPts}$ pri DBSCAN določimo empirično ali s pomočjo analitičnih metod, saj močno vplivata na obliko in število dobljenih gruč. Parameter $\mathrm{minPts}$, ki določa minimalno število točk v okolici, naj bo vsaj $d + 1$, kjer je $d$ dimenzionalnost podatkov, kar zagotavlja, da gruča ni definirana na podlagi ene same točke. Praktično se pogosto uporabljajo vrednosti med 4 in 10, pri čemer višje vrednosti pomagajo zmanjšati vpliv šuma in osamelcev. Premajhna vrednost lahko povzroči razbitje podatkov na preveč majhnih gručah, prevelika vrednost pa lahko združi ločene gruče ali povzroči, da se nekatere manjše gruče ne prepoznajo.

Izbira optimalne vrednosti $\varepsilon$ je nekoliko zahtevnejša, saj je odvisna od porazdelitve podatkov. Ena najbolj uveljavljenih metod je analiza k-razdaljnega grafa, kjer za vsako točko izračunamo razdaljo do njenega $k$-tega najbližjega soseda (običajno je $k = \mathrm{minPts}$), te razdalje uredimo naraščajoče in narišemo graf. Prelomna točka (*elbow point*), kjer razdalje začnejo hitro naraščati, predstavlja dobro izbiro za $\varepsilon$. Alternativno lahko različne vrednosti $\varepsilon$ preizkušamo eksperimentalno in opazujemo, kako se gruče oblikujejo. Premajhna vrednost vodi v številne

majhne gruče in veliko osamelcev, prevelika vrednost pa lahko združi ločene gruče v eno samo.

Dodatna možnost za oceno kakovosti gručenja je uporaba metričnih meril, kot sta Silhuetni koeficient ali Davies-Bouldin indeks, ki omogočata kvantitativno primerjavo rezultatov pri različnih nastavitvah parametrov. Pravilna izbira $\varepsilon$ in $\mathrm{minPts}$ je še posebej pomembna pri podatkih z neenakomerno gostoto, kjer lahko različni deli podatkovnega prostora zahtevajo različne vrednosti $\varepsilon$, kar predstavlja omejitev metode DBSCAN. V takih primerih se lahko uporabi razširitev DBSCAN, kot je OPTICS, ki omogoča prilagajanje kriterija gostote znotraj podatkovne množice.

## Gručenje v omrežjih

Gručenje v omrežjih temelji na razdelitvi vozlišč v skupine, imenovane skupnosti, kjer so vozlišča znotraj iste skupnosti močneje povezana kot z vozlišči v drugih skupnostih. Ta pristop se pogosto uporablja pri analizi socialnih omrežij, bioloških mrež in spletnih grafov, kjer tradicionalne metode gručenja, kot sta K-means ali DBSCAN, niso neposredno uporabne zaradi grafovske narave podatkov. Če podatki prvotno niso podani kot omrežje, jih lahko pretvorimo v graf tako, da definiramo vozlišča kot podatkovne primere in povezave kot odnose med njimi, pri čemer se uteži povezav določijo na podlagi podobnosti med podatkovnimi točkami. Pogosti načini konstruiranja grafov iz podatkov vključujejo k-najbližje sosede (k-NN graf), epsilon-sosedstva, ali graf podobnosti na podlagi korelacije.

Ena najpreprostejših metod za gručenje v omrežjih je razširjanje oznak (Label Propagation Algorithm, LPA), ki izkorišča strukturo povezav v grafu za dodelitev vozlišč v skupnosti. Algoritem začne z naključnimi oznakami skupnosti, nato pa vsako vozlišče posodobi svojo oznako na podlagi najpogostejše oznake svojih sosedov. Ta proces se ponavlja iterativno, dokler oznake ne postanejo stabilne ali se doseže maksimalno število iteracij. Ker LPA ne zahteva vnaprej določenega števila gruč in deluje linearno glede na število povezav, je zelo učinkovit pri velikih omrežjih. Vendar pa zaradi svoje naključne inicializacije lahko različne izvedbe algoritma privedejo do nekoliko različnih rezultatov.

Glavne prednosti metode razširjanja oznak so njena računska učinkovitost in sposobnost zaznavanja naravnih skupnosti brez vnaprejšnjih predpostavk o njihovem številu ali obliki. Slabost pa je, da lahko pri šibko povezanih grafih ali pri grafih z neenakomerno gostoto povezav pride do nekonsistentnih rezultatov. Poleg tega LPA v osnovni obliki ne omogoča zaznavanja prekrivajočih se skupnosti, kjer lahko posamezno vozlišče pripada več

skupnostim hkrati. Za bolj robustne rezultate se pogosto uporablja v kombinaciji z drugimi metodami, kot so modularnostna optimizacija (Louvain metoda) ali spektralno gručenje na grafih, ki upoštevajo globalne lastnosti omrežja pri določanju skupnosti.

## Metoda Louvain

Louvain metoda je ena izmed najbolj uveljavljenih metod za gručenje v omrežjih, kjer vozlišča združujemo v skupnosti na podlagi optimizacije modularnosti – mere, ki ocenjuje kakovost razdelitve omrežja v skupnosti. Algoritem je hiter in učinkovit, saj uporablja hierarhično združevanje, kar mu omogoča obdelavo velikih omrežij z milijoni vozlišč. Ime je dobil po Université catholique de Louvain, belgijski univerzi, kjer je bil razvit leta 2008. Louvain metoda se uporablja na različnih področjih, kot so analiza družbenih omrežij, biologija (genske in proteinske mreže), ekonomija in optimizacija transportnih sistemov.

Metoda Louvain deluje v dveh glavnih fazah, ki se ponavljata iterativno, dokler se struktura omrežja ne stabilizira. V prvi fazi vsako vozlišče začne kot samostojna skupnost. Nato se vsako vozlišče posamično premakne v skupnost svojega soseda, če ta premik poveča modularnost omrežja, kar pomeni, da so povezave med vozlišči znotraj skupnosti gostejše, kot bi jih pričakovali pri naključni razporeditvi povezav. Postopek se ponavlja, dokler noben premik več ne izboljšuje modularnosti. Ko je ta faza zaključena, sledi druga faza, kjer se celotne skupnosti obravnavajo kot nova vozlišča in graf se zmanjša, tako da povezave med starimi skupnostmi postanejo nove povezave med združenimi vozlišči. Nato se algoritem ponovno zažene na tej zmanjšani različici omrežja in iterativno ponavlja postopek, dokler modularnost ne doseže maksimuma ali se spremembe ustavijo.

Modularnost omrežja je mera, ki ocenjuje kakovost delitve omrežja na skupnosti. Opisuje razliko med dejanskim številom povezav znotraj skupnosti in pričakovanim številom povezav, če bi bile povezave naključno razporejene po omrežju. Matematično je modularnost definirana kot:

$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

kjer je $A_{ij}$ matrika sosednosti, $k_i$ in $k_j$ sta stopnji vozlišč $i$ in $j$, $m$ je skupno število povezav, $c_i$ je oznaka skupnosti vozlišča $i$, in $\delta(c_i, c_j)$ je indikatorna funkcija, ki je 1, če sta vozlišči v isti skupnosti, sicer 0. Modularnost meri, koliko je gručenje boljše od naključne razporeditve in

se tipično giblje med 0 in 1, kjer višje vrednosti pomenijo bolj izrazite skupnosti.

Glavne prednosti Louvain metode so njena računska učinkovitost in sposobnost samodejnega določanja števila skupnosti brez potrebe po vnaprejšnjih parametrih. Ker gradi hierarhično strukturo, omogoča večnivojsko analizo skupnosti, kar je uporabno pri kompleksnih omrežjih. Slabost metode je, da rezultati niso vedno stabilni, saj lahko majhne spremembe v podatkih vplivajo na končno razdelitev skupnosti. Prav tako je metoda optimizirana za modularnost, ki pa ni vedno najboljša mera za zaznavanje skupnosti v nekaterih vrstah omrežij, kot so zelo redko povezane mreže ali mreže z dinamičnimi spremembami.

## Mešanica Gaussovih porazdelitev

Mešanica Gaussovih porazdelitev (angl. Gaussian Mixture Models, GMM) je verjetnostni pristop k gručenju, kjer predpostavimo, da so podatki generirani iz več Gaussovih porazdelitev, pri čemer vsaka porazdelitev ustreza eni gruči. Namesto da bi vsak podatkovni primer pripadal natanko eni gruči, GMM določi verjetnost pripadnosti vsaki gruči, kar omogoča mehko gručenje (angl. *soft clustering*).

Matematično modeliramo podatkovno množico kot kombinacijo $k$ normalnih porazdelitev:

$$p(\mathbf{x}) = \sum_{i=1}^{k} \pi_i \cdot \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$$

kjer je $\pi_i$ utež $i$-te komponente oziroma pripadnost primera $i$-ti gruči (z verjetnostjo $\sum \pi_i = 1$), $\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$ pa Gaussova porazdelitev s srednjo vrednostjo $\mu_i$ in kovariančno matriko $\Sigma_i$ .

Algoritem GMM temelji na pričakovalno-maksimizacijskem (Expectation-Maximization, EM) postopku:

1. Pričakovalni korak (E-step): Za vsak podatkovni primer izračunamo verjetnost pripadnosti vsaki gruči glede na trenutno oceno parametrov.
2. Maksimizacijski korak (M-step): Posodobimo parametre (srednje vrednosti $\mu_i$, kovariančne matrike $\Sigma_i$ in uteži $\pi_i$) tako, da maksimiziramo verjetnost opazovanih podatkov.

3. Postopek ponavljamo, dokler parametri ne konvergirajo.

GMM je fleksibilnejši od postopka voditeljev, saj omogoča gruče različnih oblik in velikosti zaradi uporabo poljubnih kovariančnih matrik. Omogoča tudi mehko pripadnost primerov h gruči, kar je uporabno pri podatkih z nejasnimi mejami. Glavna slabost metode je, da zahteva izbiro števila komponent $k$, kar lahko določimo s kriteriji, kot sta Bayesov informacijski kriterij (BIC) ali Akaikejev informacijski kriterij (AIC). Poleg tega je GMM računsko zahtevnejši od postopka voditeljev, saj vključuje verjetnostne izračune in obravnava polne kovariančne matrike.

# Ocenjevanje kakovosti gručenja

Metode gručenja so nenadzorovane, kar pomeni, da pri njihovi uporabi običajno nimamo znanih resničnih oznak podatkov. Zato je ključnega pomena, da ocenimo kakovost dobljenih gruč in preverimo, ali so identificirane skupine smiselne ter skladne s strukturo podatkov. Brez ustreznih metrik lahko algoritmi vrnejo gruče, ki so računsko optimalne, a nimajo pravega pomena v kontekstu podatkov. Dobra ocena gručenja omogoča primerjavo različnih metod, izbiro optimalnega števila gruč ter identifikacijo morebitnih slabosti v modelu, kot so prekomerno razdrobljene ali nepravilno združene gruče.

Pri ocenjevanju kakovosti gručenja uporabljamo različne pristope. Notranje mere ocenjujejo gruče zgolj na podlagi značilnosti podatkov, pri čemer merijo gostoto znotraj gruč in razdaljo med njimi. Tipične notranje mere so silhuetni koeficient, Dunnov indeks in razmerje znotrajgručne in medgručne variance (Davies-Bouldin indeks). Zunanje mere, kot sta Randov indeks in F-mikro povprečje, zahtevajo primerjavo z znanimi oznakami podatkov in se uporabljajo, kadar imamo referenčno razvrstitev. Te mere pravzaprav merijo ujemanje med dvema razvrstitvama.

Obstajajo tudi stabilnostne mere, ki preverjajo, ali so dobljene gruče konsistentne pri različnih podmnožicah podatkov.

## Silhuetna metoda

Med notranjimi merami je, sicer nekako presenetljivo zaradi njene enostavnosti, ena najpogosteje uporabljenih in najbolj intuitivnih silhuetna metoda, ki omogoča oceno kakovosti gručenja brez potrebe po zunanjih oznakah. Silhuetni koeficient ocenjuje, kako dobro je

posamezna točka uvrščena v svojo gručo glede na bližino do drugih gruč. Za vsako podatkovno točko $i$ se najprej izračuna povprečna razdalja do vseh drugih točk znotraj iste gruče ($a(i)$, notranja kohezija). Nato se izračuna povprečna razdalja do vseh točk v najbližji sosednji gruči ($b(i)$, zunanja separacija). Silhuetni koeficient točke je nato definiran kot:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Vrednost $s(i)$ je med -1 in 1, kjer vrednosti blizu 1 pomenijo, da je točka jasno ločena od drugih gruč in dobro pripada svoji gruči, vrednosti blizu 0 pomenijo, da leži na meji med gručami, vrednosti pod 0 pa kažejo, da je točka napačno uvrščena. Povprečna vrednost silhuetnega koeficienta za celoten nabor podatkov je dober kazalnik kakovosti gručenja – višje vrednosti pomenijo boljše gručenje. Ta metoda se pogosto uporablja za določitev optimalnega števila gruč, saj lahko s primerjavo povprečnega silhuetnega koeficienta za različne vrednosti $k$ izberemo tisto, kjer je gručenje najbolj izrazito. Silhuetna metoda je posebej uporabna pri ocenjevanju metod, kot sta K-means in Gaussian Mixture Models, vendar je manj primerna za metode, ki ne temeljijo na razdaljah, kot so nekatere grafovske metode gručenja.

## Indeks po Randu

Randov indeks (Rand Index, RI) in prilagojeni Randov indeks (Adjusted Rand Index, ARI) sta zunanji metri za ocenjevanje kakovosti gručenja, kar pomeni, da ju uporabljamo, kadar imamo referenčne oznake (resnične skupine), s katerimi lahko primerjamo dobljene gruče. Oba merita, kako dobro se gručenje ujema z dejansko razdelitvijo podatkov.

Randov indeks temelji na številu parov podatkovnih točk, ki so bodisi pravilno razvrščeni v isto gručo bodisi pravilno ločeni v različne gruče. Če imamo n podatkovnih primerov, obstaja $\binom{n}{2}$ možnih parov podatkovnih točk, ki jih lahko primerjamo. Randov indeks se izračuna kot:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

kjer:

- TP (True Positives): Pari točk, ki so v isti gruči tako v resnični kot v predvideni razvrstitvi.

- TN (True Negatives): Pari točk, ki so v različnih gručah v obeh razvrstitvah.
- FP (False Positives): Pari točk, ki so v isti gruči v predvideni razvrstitvi, a v različnih v resnični.
- FN (False Negatives): Pari točk, ki so v različnih gručah v predvideni razvrstitvi, a v isti v resnični.

Randov indeks zavzame vrednosti med 0 in 1, kjer je 1 popolno ujemanje gručenja s pričakovano razvrstitvijo, 0 pa pomeni popolno neskladje.

Randov indeks ima težavo, da lahko pri naključnem gručenju še vedno vrne visoke vrednosti zaradi srečnega ujemanja parov. Da bi to odpravili, se uporablja prilagojeni Randov indeks (Adjusted Rand Index, ARI), ki normalizira vrednosti, tako da upošteva pričakovano vrednost naključnega gručenja. ARI je definiran kot:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

kjer je $E[RI]$ pričakovana vrednost Randovega indeksa, če bi gručenje bilo naključno. Prilagojeni Randov indeks se lahko giblje od -1 do 1, kjer:

- 1 pomeni popolno ujemanje,
- 0 pomeni naključno gručenje,
- negativne vrednosti pomenijo slabše od naključja.

Pri izračunu prilagojenega Randovega indeksa (ARI) je pričakovana vrednost Randovega indeksa $E[RI]$ ključna za normalizacijo, da se upošteva, kakšno vrednost bi imel Randov indeks, če bi gručenje potekalo naključno. Naj bo:

- $n$ število podatkovnih primerov,
- $n_{ij}$ število skupnih točk med gručo $i$ v resnični razvrstitvi in gručo $j$ v predvideni razvrstitvi,
- $a_i$ število točk v gruči $i$ v resnični razvrstitvi,
- $b_j$ število točk v gruči $j$ v predvideni razvrstitvi.

Pričakovana vrednost Randovega indeksa se izračuna na podlagi kombinatorike parov točk, pri čemer se upošteva, da se število pravih pozitivnih (TP) in pravih negativnih (TN) pri naključnem gručenju porazdeli glede na velikost gruč. Zapišemo pričakovano vrednost kot:

$$E[RI] = \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}$$

kjer:

- $\binom{a_i}{2} = \frac{a_i(a_i-1)}{2}$ predstavlja število parov znotraj gruče $i$ v resnični razvrstitvi,
- $\binom{b_j}{2} = \frac{b_j(b_j-1)}{2}$ predstavlja število parov znotraj gruče $j$ v predvideni razvrstitvi,
- $\binom{n}{2} = \frac{n(n-1)}{2}$ je skupno število možnih parov v podatkovni množici.

Zaradi te normalizacije je prilagojeni Randov indeks (ARI) definiran kot:

$$ARI = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - E[RI]}{\frac{1}{2}\left(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right) - E[RI]}$$

Tako ARI odpravlja pristranskost Randovega indeksa in omogoča bolj pošteno oceno kakovosti gručenja, saj daje vrednost 0, kadar je gručenje naključno, in 1, kadar se popolnoma ujema z referenčno razvrstitvijo.

Ključna razlika med RI in ARI je, da RI ni normaliziran in lahko pri naključnem gručenju vrne visoke vrednosti, medtem ko ARI popravi to pomanjkljivost z upoštevanjem pričakovane porazdelitve naključnih rezultatov. Zaradi tega se ARI pogosteje uporablja v praksi, saj daje bolj realno sliko kakovosti gručenja.

# Razlaga dobljenih gruč

Gručenje podatkov omogoča odkrivanje skritih vzorcev, a brez ustrezne interpretacije dobljenih skupin je njihova uporabnost omejena. Razumevanje, kaj posamezna gruča pomeni, je ključno za sprejemanje odločitev, oblikovanje strategij in nadaljnjo analizo podatkov. Na primer, v medicini gruče lahko označujejo različne podtipe bolezni, v trženju različne segmente strank, v genomiki pa funkcionalno povezane gene. Brez razlage pa je tveganje napačne uporabe rezultatov veliko, saj se lahko gruče oblikujejo na podlagi nerelevantnih značilk ali statističnih artefaktov. Zato je pomembno uporabiti metode, ki omogočajo vpogled v značilnosti in strukturo dobljenih skupin. Spodaj naštejemo

## Povprečne vrednosti in značilni predstavniki

Ena najpreprostejših metod za interpretacijo gručenja je pregled povprečnih vrednosti značilk

v posamezni gruči, kar omogoča hitro razumevanje glavnih razlik med skupinami. To tehniko lahko najbolje uporabljamo v povezavi s tehniko rangiranja značilk pomembnih za gručenje, saj je ideja, da primerjamo le vrednosti tistih značilk, ki so med sabo najbolj različne med skupinami. Alternativno lahko določimo najbolj reprezentativne točke gruče, kot so medoidi v K-medoids ali podatkovne točke, ki so najbližje središču gruče.

## Vizualizacija podatkov

Pri večdimenzijskih podatkih bi bilo koristno zmanjšati dimenzionalnost in vizualizirati gruče. Glavne tehnike vključujejo metodo glavnih komponent, ki omogoča projekcijo podatkov v 2D ali 3D prostor, t-SNE (t-distributed Stochastic Neighbor Embedding), ki poudarja lokalne vzorce v podatkih, in UMAP (Uniform Manifold Approximation and Projection), ki je podobna t-SNE, a ohranja globalno strukturo podatkov. O teh metodah bomo podrobno spregovorili v naslednjem poglavju.

## Identifikacija značilk pomembnih za gručenje

Za vsako značilko lahko ocenimo njen prispevek k oblikovanju gruč. To lahko storimo s testiranjem statističnih razlik (npr. ANOVA za numerične podatke, hi-kvadrat test za kategorične podatke) ali s tehnikami poudarjanja značilk, kot so SHAP vrednosti, ki omogočajo oceno vpliva posameznih spremenljivk na gručenje.

## Semantična interpretacija gruč

V nekaterih domenah, kot so obdelava naravnega jezika in biomedicina, lahko gruče interpretiramo s pomočjo ontologij in semantičnih modelov. Na primer, gruče v besedilnih podatkih lahko analiziramo z metodo TF-IDF za ključne besede, v bioloških podatkih pa s pomočjo Gene Ontology, ki povezuje gruče genov s funkcionalnimi kategorijami.

## Pravila in odločitvena drevesa za razlago gruč

Za lažjo razlago gruč lahko zgradimo odločitvena drevesa ali pravila na podlagi značilk, ki najbolj ločujejo gruče. Ta pristop omogoča generiranje preprostih pravil, kot je npr. "*Če je starost > 50 in krvni tlak > 140, potem je verjetnost pripadnosti gruči X 85%*". To olajša razumevanje kompleksnih rezultatov in omogoča njihovo uporabo v realnih aplikacijah.

Razlaga gručenja je ključna za smiselno uporabo rezultatov, saj brez nje dobljene skupine

ostanejo zgolj matematične entitete brez jasnega pomena. Z uporabo različnih metod za vizualizacijo, analizo značilk, semantično razlago in generiranje pravil lahko pridobimo boljše vpoglede v strukturo podatkov in s tem izboljšamo uporabnost modelov gručenja v različnih aplikacijah.