Klasterinė analizė

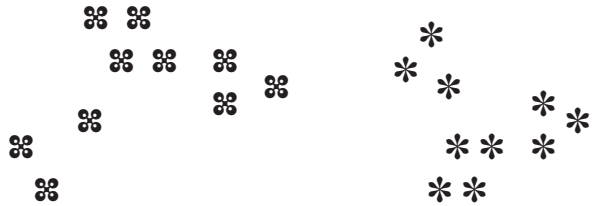
Klasterio sąvoka

Taikydami klasterinę analizę, remiantis turimais duomenimis, nustatome objektų panašumą ir suskirstome juos į *klasterius*.

- Klasteris - panašių objektų grupė.
- Objektų panašumas nusakomas skaitiniais panašumo matais.
- Skirstydami objektus į klasterius, dažniausiai net nežinome, kiek klasterių turimoje duomenų aibėje realiai egzistuoja (ir ar išvis egzistuoja).
- Klasterinėje analizėje imtis neturi klasės kintamojo, pagal kurio reikšmes būtų galima patikrinti ("kontroliuoti") ar žinomas mokymo imties įrašas pateko į "teisingą" klasterį.

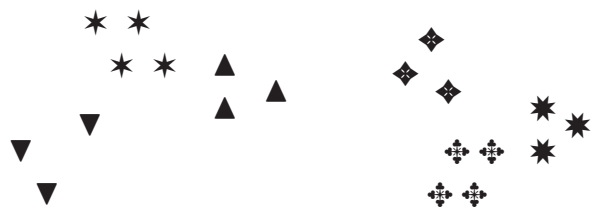
Taškų aibės klasterizavimo pavyzdys

(a) Visi taškai



(b) Du klasteriai

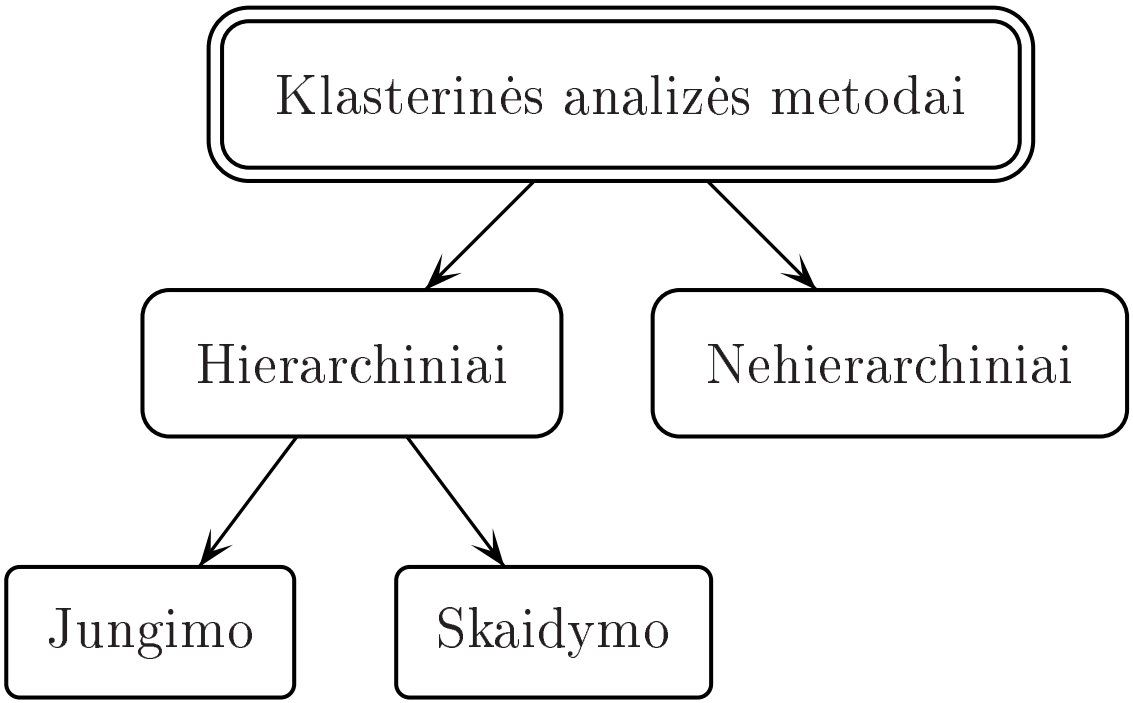
(c) Keturi klasteriai



(d) Šeši klasteriai

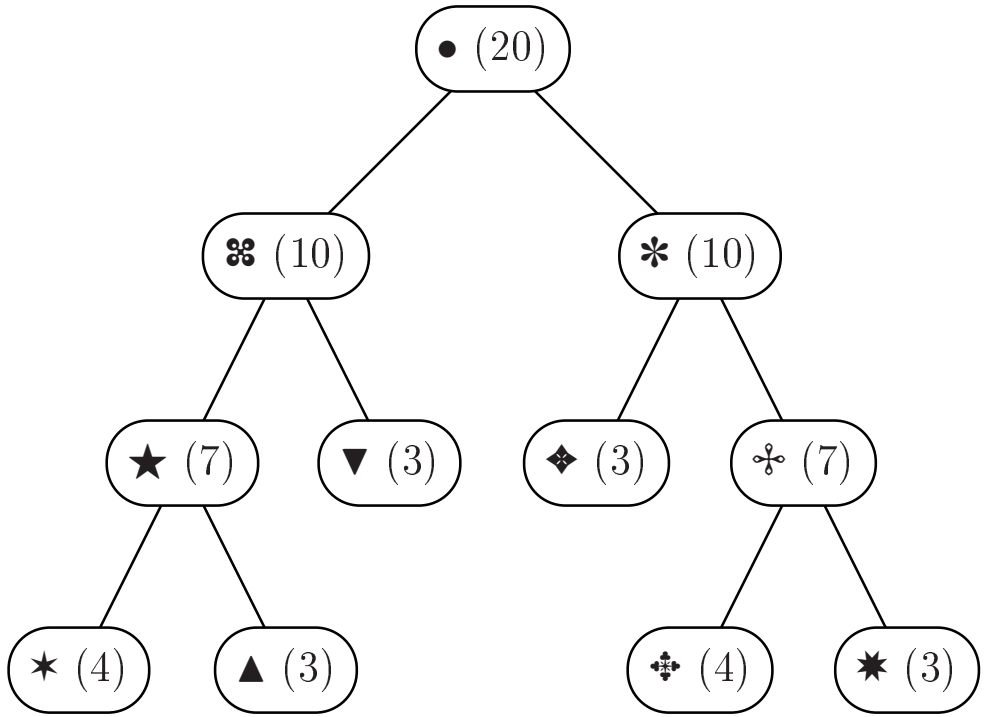
Klasterinės analizės metodų klasifikacija

Dažnai klasterių sudarymo metodai skirstomi pagal galimą klasterių tarpusavio išsidėstymą.



Hierarchiniai metodai

Visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos savo ruožtu dar mažesni ir t.t.



Čia kiekvienas klasteris (medžio viršūnė) vaizduojamas jį sudarančių taškų žyme ir taškų skaičiumi.

Hierarchiniai metodai

Priklausomai nuo to kaip "auginamas" klasterių medis, hierarchiniai metodai skirstomi į jungimo ir skaidymo metodus.

- Jungimo metodai konstruoja medį nuo lapų, t.y. smulkius klasterius jungia vis į stambesnius, kol galų gale lieka vienas.
- Skaidymo metodai "augina" medį nuo šaknies - vienintelį klasterį nuosekliai skaido į dalis.

Nehierarchiniai metodai

Nehierarchiniai metodai tiesiog skaido turimą duomenų aibę į K nesikertančių poaibių (klasterių) taip, kad artimi objektai patektų į vieną poaibį.

Tokie metodai paprastai taikomi tada, kai iš anksto žinomas (pasirenkamas) klasterių skaičius K.

Jungimo metodai

Tegul $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ - duomenų (objektų) aibė, $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ - atstumas tarp objektų $\mathbf{x}_i$ ir $\mathbf{x}_j$.
Atstumai sudaro simetrinę *atstumų matricą* $(d_{ij})_{N\times N}$.
Hierarchiniai jungimo metodai, pradedant pavieniais objektais, nuosekliai jungia du artimiausius klasterius, kol lieka tik vienas klasteris. Tokio algoritmo schema

1. Turime N klasterių po 1 objektą
2. Apskaičiuojame atstumų matricą $(d_{ij})_{N\times N}$
3. **repeat**
4. Nustatome du artimiausius klasterius U ir V
5. U ir V sujungiami į vieną klasterį $U \cup V$
6. Transformuojama atstumų matrica
7. **until** Lieka tik vienas klasteris

Atstumai tarp klasterių

Dažniausiai naudojami atstumai $d(U, V)$ tarp klasterių $U \subset X$ ir $V \subset X$

Atstumas	$d(U, V)$ formulė
Artimiausio kaimyno	$d(U, V) = \min\limits_{\mathbf{x} \in U,\, \mathbf{x}' \in V} d(\mathbf{x}, \mathbf{x}')$
Tolimiausio kaimyno	$d(U, V) = \max\limits_{\mathbf{x} \in U,\, \mathbf{x}' \in V} d(\mathbf{x}, \mathbf{x}')$
Vidutinis	$d(U, V) = \dfrac{1}{\lvert U \rvert \cdot \lvert V \rvert} \sum\limits_{\mathbf{x} \in U} \sum\limits_{\mathbf{x}' \in V} d(\mathbf{x}, \mathbf{x}')$
Centrų	$d(U, V) = d(\mathbf{x}_U, \mathbf{x}_V)$, $\mathbf{x}_U, \mathbf{x}_V$ - klasterių U ir V centrai

Klasterių centrai

Klasterio centro sąvoka priklauso nuo objektų atstumo apibrėžimo. Pastebėsime, kad $\mathbf{x}_U$ ne visada yra klasterio U objektas. Pavyzdžiui, kai $X \subset \mathbb{R}^p$, o atstumas - Euklido metrika $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(x_{i1} - x_{j1})^2 + ... + (x_{ip} - x_{jp})^2}$, centru dažniausiai laikomas "vidutinis" klasterio U objektas

$$\mathbf{x}_U = \frac{1}{|U|} \sum_{\mathbf{x} \in U} \mathbf{x}\,, \quad |U| \; - \text{ klasterio } U \text{ objektų skaičius.}$$

Šiuo atveju, gali būti skaičiuojamas ir vadinamasis Ward'o atstumas

$$d_W(U, V) = \frac{|U| \cdot |V|}{|U| + |V|} \|\mathbf{x}_U - \mathbf{x}_V\|^2 .$$

Jam būdinga tai, kad sujungus du artimiausius klasterius, visada yra minimizuojama objektų atstumų nuo jų klasterių centrų kvadratų suma.

K - vidurkių metodas

Didesnės objektų aibės dažnai klasterizuojamos nehierarchiniais metodais. Vienas iš tokių ir yra K-vidurkių metodas. Jis plačiai taikomas ir kartu gana paprastas.
Trumpas algoritmo pseudokodas

1. Pasirekame K pradinių klasterių centrų
2. **repeat**
3. Objektai suskirstomi į K klasterių, kiekvieną objektą priskiriant artimiausiam centrui
4. Perskaičiuojami klasterių centrai
5. **until** Klasterių centrai nebekinta

Konkreti algoritmo realizacija priklauso nuo turimų duomenų tipo ir naudojamo objektų atstumo mato. Nuo to priklauso ir klasterio centro sąvoka.

K - vidurkių metodas

Tegu klasterizuojami objektai yra Euklido erdvės vektoriai, t.y. $X \subset \mathbb{R}^p$, $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$.
Kai objektų aibė X skaidoma į K nesikertančių klasterių

$$X = C_1 \cup C_2 \cup \cdots \cup C_K \,,$$

Natūralu manyti, kad esant gerai klasterizacijai, objektai turi būti kuo arčiau klasterių centrų. Vadinasi tikslo funkcija, kurią reikia minimizuoti, parenkant klasterių centrus, yra

$$SSE = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{x}_{C_i}\|^2 \,.$$

Galima įrodyti, kad SSE įgyja mažiausią reikšmę, kai $\mathbf{x}_{C_i}$ yra lygus klasterio C_i vektorių vidurkiui.

Pradinių centrų parinkimo problema

K-vidurkių metodo trūkumai

- Klasterių skaičių reikia nustatyti iš anksto. Tuo pačiu tarsi primetama tam tikra duomenų struktūra, nebūtinai sutampanti su objektyviai egzistuojančia.
- Pradinių K centrų parinkimo problema. Pavyzdžiui, būtina atsižvelgti į galimas išskirtis duomenų aibėje, stengtis kad pradiniai centrai nepriklausytų vienam klasteriui.

Naudojamos įvairios pradinių centrų parinkimo strategijos: nuo atsitiktinio parinkimo iki preliminarios hierarchinės klasterizacijos į K klasterių .