**Asier Gonzalez Zabala eta Igor Ruiz Agúndez**
DeustoTech Computing, Tecnológico Fundación Deusto

# **Kluster analisia**, multzokatze tekniken gakoak

## Laburpena

Zer da kluster bat? Zergatik erabiltzen dira klustering teknikak? Zertarako erabiltzen dira? Zein arazori egin behar diote aurre kluster tekniken erabiltzaileek? Zein urrats eman behar dira klusterizazio bat egiteko?

Artikulu honetan, galdera hauei erantzuten saiatuko gara, eta, horretarako, kluster analisiaren munduaren oinarrizko kontzeptuak aurkeztuko ditugu.

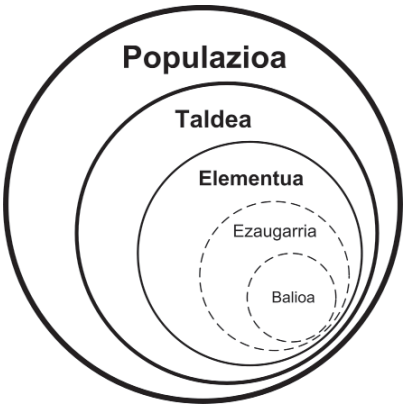
## Definizioa eta funtzionamendua

Kluster analisiak informazioa taldekatzea du helburu. Behatokiak, bektoreak eta bestelako datuak multzokatzen ditu, eta, horrela, euren egitura naturala aurkitzen du.

Definizio formalagoek kluster analisia sailkapen estatistikoa erabiltzen duen teknika bezala aurkezten dute. Populazio jakin bateko datuen ezaugarrien konparaketa kuantitatiboa egiten du, eta, horrela, teknika horrek datuak taldeetan sailkatzen ditu.

Analisi honek sorturiko taldeak (klusterrak) esanguratsuak eta erabilgarriak dira; izan ere, klusterrak erabilitako datuen egitura naturala islatzen dute. Egitura horri esker, datuen bidez, ezagutza geureganatzen dugu, eta, horrela, hainbat arlotan (psikologian, biologian, estatistikan, ikaskuntza automatikoan) erabilgarria izan daitekeen ezagutza lortzen dugu.

Taldekatzea egiteko, datuen barne informazioa erabili behar dugu; alegia, datuak deskribatzen dituzten balioen bidez eta barne erlazioen azterketaren bidez sortuko ditugu taldeok. Helburua da datuen barneko instantzia (balio) bakoitza pareko talde batean sailkatzea eta, aldi berean, talde barneko instantziak euren artean erlazionatuta egotea. Gainera, beste taldeak ez bezalakoak izango dira, eta, horrela, talde bakoitzaren adierazgarritasuna bideratuko dute.




**1. ilustrazioa.** Kluster baten osagaiak

Modu horretan, klusterizazio teknika gehienek datuak taldeetan biltzen dituzte, baina datu bakoitza batean baino ez da biltzen, taldeen arteko talkarik eragin gabe. Horrela, zenbat eta talde barnean antzekotasun handiagoa eta beste taldeekiko antzekotasun txikiagoa egon, klusterizazioa orduan eta esanguratsuagoa izango da.

Klusterizazioak instantzia bakoitzari etiketa bat esleituko dio, klusterraren identifikadorea izango dena. Barne funtzionamendu hau dela eta, batzuetan, klusterizazioa sailkapen ez-ikuskatu moduan ere ezagutzen da. Klusterizazioak ez du aurretik zehaztutako etiketarik erabiltzen, balidazioak egiteko salbu.

## Zergatik erabili klustering teknikak?

Klustering analisiak hainbat diziplinatan dauka aplikazioa. Lan berri asko teknika honetan oinarrituta daude eta, ondorioz, oso zaila da zenbat aplikaziok edo zientzia arlok erabiltzen dituzten esatea. Adibidez, irudien segmentazioan (ordenagailu bidezko ikuspenetan oinarritutako ohiko arazo bat), klustering teknikak erabiltzen dira. Informazioa arin berreskuratzeko, dokumentuak taldekatu egiten dira, eta, horrela, topikoen hierarkia berria sortzen da.

Bestalde, klustering teknikak beste hainbat erabilera ditu. Esate baterako, marketinean, produktuak egoki merkaturatzeko, bezeroak teknika horren bitartez multzokatzen dira, eta langileak egoki administratzeko eta planifikatzeko ere balio du. Biologian ere aplikazioa dauka: zientzialariek genoman datuak sailkatzeko eta ikertzeko erabiltzen dute.

Aurreko adibideak albo batera utzita, etiketak, normalean, hurrengo helburuak betetzeko erabili ohi dira:

- **Azpiko egitura**: hipotesiak sortzeko, arazoak detektatzeko eta ezaugarri aipagarrienak identifikatzeko.

**2. ilustrazioa.** Kluster analisiak elementuak taldekatzen ditu. (Argazkilaria Brent Danley, license cc attribution noncommercial share alike)

Laburbilduz, hurrengo arloetan erabiltzen da klustering teknika:

- Ikaste automatikoa.
- Adimen Artifiziala.
- Patroien azterketa.
- Ingeniaritza Mekanikoa, Elektrikoa, eta bestelakoak.
- *Web*-meatzaritza.
- Datu espazialen analisia.
- Testu dokumentuen bilduma.
- Irudien segmentazioa.
- Medikuntza, Biologia, Mikrobiologia, Paleontologia, Psikiatria, Soziologia, Psikologia, Arkeologia, Hezkuntza, Marketina.

Ikusi dugunez, hainbat arlotan erabil daitezke klustering teknikak.

**Zein urrats eman behar ditu klusterizazio batek?**

Klusterizazio prozesu orok, behintzat, honako urratsak eman behar ditu:

- Populazioen errepresentazioa: populazioen instantziak deskribatzen dituzten balioak hautatu eta txukundu.
- Algoritmoaren diseinua edo aukeraketa: existitzen diren algoritmoetatik egokiena aukeratu edo eta beste bat sortu.
- Modelizazioa: algoritmoaren exekuzioaren ondorioz lortzen den emaitza; hau da, etiketak jarrita.
- Optimizazioa: emaitzak hobetzen saiatu, eta, horretarako, ahalik eta kalitate handiena bilatu eta beharrezkoak diren bitartekoak murriztu.
- Balidazioa: eskuraturiko patroiaren balioa egiaztatu. Lortutako emaitzek errealitatea erakusten dutela frogatu.
- Emaitzen interpretazioa: modelizatu, optimizatu eta balioztatutako emaitzetatik erabilgarria izango den ezagutza jaso.

**Esker onak**

- Larraitz Uriarte eta Miren Agurtzane Mallona eskertu nahi ditugu artikuluaren zuzenketan eginiko lanarengatik.

**3. ilustrazioa.** Klustering teknikek edozein elementu taldeka dezakete. Adibidez, ilustrazioko pilotak kolorearen arabera bildu dezakete. (Argazkilaria fontplaydotcom, license cc attribution

- **Natura sailkapena**: forma edo organismoen artean antzekotasun maila lortzeko (harreman filogenetikoa).
- **Konpresioa**: klustering prototipoen bitartez, datuak antolatzeko eta laburbiltzeko.

## Zertarako erabiltzen da?

Klustering teknikak hainbat arlotan erabiltzen dira. Datuak ulertzeko orduan, klusterrak etiketa moduan erabiltzen dira eta kluster analisiak etiketak automatikoki aurkitzeko balio du. Hona hemen adibide batzuk:

- **Biologia**. Biologoek hainbat urte igaro dituzte izaki bizien taxonomia sortzen: phylum, klaseak, ordena, familia, generoa eta espeziea sailkatzen. Horrela, ez da harritzekoa arlo honetako lehenengo sailkapenlanak kluster analisia erabili izana, taxonomia matematiko bat sortzeko eta, ondorioz, egitura biziak sailkatzeko. Gaur egun, klustering teknika kopuru handiko informazio genetikoa analizatzeko erabiltzen da. Adibidez, genetikan, antzeko funtzionaltasunak dituzten taldeak bilatzeko erabili ohi da.
- **Informazioa eskuratzea**. Internet-en mila milioi *web* orri daude eta bilaketa baten emaitza milaka orrikoa izan daiteke. Klustering teknikak bilaketa hau kluster txikietan taldekatzeko aukera ematen du. Adibidez, «pelikula» terminoa bilatzen badugu, teknika honek jasotako emaitzak analizatu eta taldeetan banatzen ditu (kritikak, trailerrak, puntuazioak,...). Era berean, talde horiek, beste talde txikiago batzuetan bana daitezke eta, horrela, egitura hierarkiko bat sortu.
- **Klima**. Lurreko klima ulertzeko, ezinbestekoa da itsaso eta atmosferan patroiak bilatzea. Helburu hori betetzeko, klustering tekniken bitartez, poloetan eta itsasoan jasotzen den presio atmosferikoa aztertzen da.
- **Psikologia eta Medikuntza**. Gaixotasun batek hainbat aldaera eduki dezake eta, analisiaren bitartez, aldaera horiek talde txikietan batu daitezke. Adibidez, depresio motak bila daitezke.
- **Datu espazialen analisia:** Gaur egun, teknologia aurreratuen bitartez (teleskopio espazialak, Informazio Geografikoko Sistemak –*GIS*–, ekipo medikoak), datu espazial asko eta asko lortzen dira. Beraz, datu horien analisia oso lan zaila eta garestia izaten da langileentzat. Arazo hori konpontzeko, klustering teknikak erabiltzen dira, eta, horretarako, ezaugarri interesgarriak identifikatu eta antzeko patroiak bilatzen ditu.
- **Negozioetan**: Klustering tekniken bitartez, saltzaileek talde adierazgarriak datu baseetan bilatu eta talde horien erosketa patroietan oinarritutako ezaugarriak eskuratzen dituzte.

**Zein ohiko arazori egin behar zaio aurre, klusterrarekin lan egitean?**

- Zein klusterizazio ezaugarri erabil daitezke?
- Datuak normalizatu ahal dira?
- Datuek balio atipikoak dituzte?
- Zelan zehazten da *pair-wise* deritzon antzekotasuna? Hau da, zein da datuen arteko antzekotasuna?
- Zenbat kluster agertzen dira datuetan?
- Zein klustering metodo erabil daiteke? Zein ezarpenekin?
- Datuek ba al dute klusterra sortzeko joerarik? Klusteriza daitezke?