Kasutamiseks kursusel „Andmeanalüüsi meetodite praktiline kaustamine sotsioloogilises uurimistöös”

Koostaja: Kadri Täht
Koostatud: sügis 2010

KLASTERANALÜÜS

Loengukonspekt

Üldine eesmärk
Mitmel juhul on uurijad silmitsi tõsiasjaga, et kuidas n-ö sisukalt/mõttekalt koondada vaadeldavaid andmeid ehk siis arendada välja taksonoomiaid/süstemaatikaid.

Mõiste klasteranalüüs (esimest korda kasutas seda Tryon, 1939) hõlmab endas rida algoritme ja meetodeid sarnaste objektide/asjade grupeerimiseks vastavasisulistesse kategooriatesse. Teisisõnu, klasteranalüüs (KA)on uurimuslik andmeanalüüsi meetod, mille eesmärgiks on sorteerida erinevad objektid gruppidesse sellisel moel, et ühte gruppi kuuludes on kahe objekti vahel maksimaalselt tugev seos ning vastupidisel juhul minimaalne seos. Seega saab KA kasutada selleks, et ‘avastada’ andmetes olevaid struktuure ilma neile seletust/interpretatsiooni andmata. Teisisõnu, KA abil lihtsalt ’avastatakse’ andmetes struktuure ilma, et seletatakse, miks need eksisteerivad.

Me tegeleme klasterdamisega tavaelu pea iga aspekti puhul. Näiteks sööklas ühte lauda jagavaid tudengeid võib käsitleda kui inimeste klastrit. Toidupoes on sarnased tooted (näiteks erinevad lihad, erinevad juurviljad) paigutatud lähestikku. Jne. Jne

Kasutusala
Klasterdamise tehnikaid on kasutatud suure hulga uurimisprobleemide puhul. Näiteks meditsiinis haiguste, ravi või haigussümptomite võib viia väga kasulike taksonoomiateni. Psühhiaatrias paranoia, skisofreenia jne sümptomite korrektne klastrite diagnoosimine on õnnestunud teraapia eelduseks.

Igal juhul kui meil on vaja klassifitseerida suurt hulka informatsiooni hõlpsasti koheldavateks sisulisteks osadeks, siis KA-st võib seejuures olla palju kasu.

Klasteranalüüsi põhimeetodid
Klaster – teatud sarnaste objektide hulk.

Klastrit iseloomustab:

- Punkide paigutuse tihedus;
- Hajuvus klastri sees;
- Klastri mõõtmed;
- Vorm;
- Kattuvus (kas näiteks klastrites leidub ühiseid elemente. Enamasti vaatleme siiski klasterid, kus ei ole kattuvaid elemente)

a) Hierarhilised meetodid

… iga järgneva rühmituse sees sisaldub eelmine; iga järgnev rühm tekib põhinedes eelmisel.

Eristatakse:

- aglomeratiivsed, so ühendavad meetodid (iga indiviid alguses üks klaster ja viimaks moodustavad ühe klastri)
- liigendavad meetodid (algul kõik indiviidid üks klaster ja siis hakatakse neid üksteisest lahku lööma)

b) Iteratiivsed e. sammuviisilised meetodid

… toimub tsüklilisena. Pidevalt toimub tulemuste parendamine ja indiviidide ümberrühmitamine. Eeldatakse, et liigendus koondub mingiks üheks liigendiks ehk tõeliselt kogumis esinevaks tunnuseks. Liigendus peaks järjest paranema. Tulemused erineva kujuga ning erineva kasutusega

c) Graafilised meetodid e. graafiteoorial põhinevad

d) Faktormeetodid

Sarnasus/erinevusmäärad

Hierarhilise meetodi puhul kasutatakse klastri moodustamisel objektide vahel olevaid erinevusi/saransusis või distantse. Sarnasused on teatud hulk reegleid, mida kasutatakse kriteeriumina üksikute ühikute grupeerimisel või eraldamisel. Need distantsid (sarnasused) võivad põhineda ühel dimensioonil või ka mitmesel dimensioonil, millest iga dimensioon esindab ühte reeglit või tingimust meie huviorbiidis olevate objektide grupeerimiseks. Näiteks kui me peaksime klasterdama kiirtoite, siis võime võtta arvesse nende kalorsust, nende hinda, subjektiivseid hinnanguid maitsele jne.

Kõige konkreetsem/otsesem viis arvutada objektide vahelisi kaugusi mitmedimensionaalses ruumis on arvutada eukledilisi kaugusi. Kui meil oleks kahe- või kolmedimensionaalne ruum, siis see mõõdab tegelikku geomeetrilist distantsi tunnuste vahel (näiteks kui me mõõdaks seda joonlauaga). Siiski on oluline meeles pidada, et ühendamise algoritm ei ’hooli’ sellest, et need distantsid, mis on ette antud on otsesed kaugused või mõned tuletatud kauguse mõõdud, mis võivad olla uurija jaoks palju kõnekamad/sisukamad. Seega, sõltub see uurijast, et valitaks õige meetod oma konkreetse uurimuse jaoks.

- Eukleidiline kaugus

See on ilmselt kõige sagedamini valitud kauguse tüüp. See on lihtsalt geomeetriline distants multidimensionaalses ruumis. Seda arvutatakse järgmiselt:

$$\text{kaugus}\ (x,y) = \{\sum_i (x_i - y_i)^2\}^{½}$$

Oluline on silmas pidada, et eukleidilist kaugust (ja ka selle ruutu) arvutatakse reeglina algandmete pealt ning mitte standardiseeritud andmete pealt. Sellel meetodil on teatud eeliseid, näiteks kahe objekti vahelise kaugus ei sõltu/muutu kolmanda objekti lisamisega analüüsi, viimane võib olla aga erand. Siiski sõltub kaugus tugevalt erinevustest skaaladel nende dimensioonide vahel, millest erinevusi arvutatakse. Näiteks kui üks dimensioonidest esindab pikkust mõõdetud sentimeetrites ja siis konverteeritakse see ümber millimeetriteks, eukleidiliste kauguste mõõdud on sellest mõjutatud (s.t. mõjutatud/kujundatud nende dimensioonide poolt, millel on suurem skaala) ja klasteranalüüsi tulemused võivad olla väga erinevad. Heaks tavaks on see, et dimensioonid muutekase skaala mõttes ühtseks.

- Eukleidilise kauguse ruut – mõnikord võib kasutada ka eukledilise kauguse ruutu, et panna kasvavalt suurem rõhk neile objektidele, mis asuvad kaugemal. Seda kaugust arvutatakse

$$\text{kaugus } (x,y) = \sum_i (x_i - y_i)^2$$

- City-block (Manhattan) distance – see kaugus on lihtsalt keskmine erinevus dimensioonide lõikes. Enamusel juhtudel see kaugusmõõt annab sarnaseid tulemusi tavalise eukleidilise kaugusega. Siiski on oluline silmas pidada, et selle mõõtmisviisi puhul üksikute suurte erinevuste (erandid) mõju on pehmendatud (kuna need ei ole ruutu võetud). Seda arvutatakse järgmiselt:

$$\text{kaugus } (x,y) = \sum_i | x_i - y_i |$$

- Chebychev distance – seda mõõtmisviisi on ehk mõttekas kasutada siis kui näiteks soovitakse kahte objekti defineerida kui ’erinevaid’ kui nad on erinevad mingis dimensioonis. Seda arvutatakse järgmiselt:

$$\text{kaugus } (x,y) = \text{Maximum } | x_i - y_i |$$

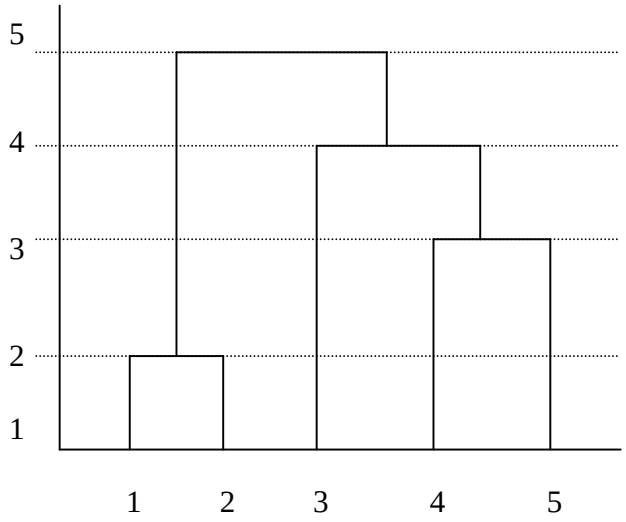
Statistilise olulisuse testimine

KA puhul räägitakse küll selle aluseks olevatest algoritmidest, kuid mitte statistilise olulisuse testidest. Tegelikult ei ole KA kuigi tüüpiline statistiline test kuna see on ’kogum’ erinevatest algoritmidest, mis ’paigutavad mingid objektid klastritesse vastavalt eelnevalt defineeritud reeglitele’. Erinevalt paljudest teistest statistilistest protseduuridest kasutatakse KA meetodit kui meil ei ole *a priori* hüpoteesi, vaid kui oleme veel oma uurimuse uurimuslikus/avastuslikus faasis. Teatud mõttes leiab KA ’olulisima võimaliku lahenduse’ Seetõttu ei ole statistilise olulisuse test siinkohal asjakohane, isegi siis kui räägitakse tõenäosustest (nagu näiteks k-keskmiste KA puhul)

Hierarhilised aglomeratiivsed rühmitusmeetodid

Liigitusmeetodis saame liigituse tulemusena liigituspuu ehk dendrogrammi. Saame ühinemiskäigu, meil võimalik valida sobiv tase sealt puust. Tekivad mitmeklastrilised lahendid ning meie valida on, mitmeklastrilise võtame.

Me alustame sellest, et iga indiviid on omaette klaster. Nüüd väikeste sammudena ’lõdvendame’ om kriteeriumeid selle kohta, et mis on unikaalne/eraldiseisev ja mis mitte. Teisisõnu, ma laseme alla seda piirangukriteeriumit mille alusel arvame kaks või enamat objekti ühte klastrisse.



Selle tulemusena *ühendame* rohkem ja rohkem objekte kokku ja agregeerime suuremaid ja suuremaid klastreid, mis sisaldavad endas üha erinevaid elemente. Lõpus, viimase sammuna on kõik objektid kokku ühendatud. Neil joonistel vertikaalse puu korral vertikaalne ja horisontaalse puu korral horisontaalne telg tähistab seotuse kaugust. Seega, iga ‘sõlme’ korral graafikul (kus uut klastrit moodustatakse) me saame välja lugeda selle distantsi kriteeriumi, mille juures need vastavad elemendid ühendati kokku uueks eraldiseisvaks klastriks. Kui andmetes on sarnaste objektide klastrite mõttes olemas selge ’struktuur’, siis seda struktuuri esitatakse hierarhiapuul kui konkreetset haru. Ühendamismeetodid eduka analüüsi tulemusena korral on võimalik eristada neid harusid ning neid ka interpreteerida.

Esimesel sammul kui iga objekt kujutab endast omaette klastrit, siis distants nende vahel on defineeritud valitud kaugusmõõdu alusel. Siiski, kui erinevad objektid on omavahel ära

ühendatud/seostatud, kuidas määratleda kaugused uute klastrite vahel? Teisisõnu, meil on vaja seose või ühendamise reeglit, mille alusel kaks klastrit on piisavalt sarnased, et need omavahel kokku siduda.

Selleks on erinevaid meetodeid. Näiteks võime ühendada kaks klastrit kokku kui igasugused kaks objekti kahes klastris on lähemal kui vastav sidumise kaugus. Teisisõne me kasutame ’lähimaid naabreid’ klastrite lõikes, et klastrite vahelisi kaugusi defineerida.

- ühe seose meetod ehk lähima naabri meetod *(single linkage method, nearest neighbour method)* Siin tegu klastrite vahelise kaugusega (varem indiviidide vahelise kaugusega), st klastrite vaheline kaugus nende lähima naabri järgi

  Algselt iga indiviid klaster, vähimate kaugustega klastrid ühendame omavahel.

  Tavaliselt ahela efekt -> üks suur ja palju väikesi klasterid. Lähima naabri meetodi puhul uut klastrit alustatakse suhteliselt raskesti, seetõttu kohati halb lähenemine. Tulemused ka kohmakamad, liiga palju indiviide.

  See meetod teatud mõttes ’seob’ objekte omavahel kokku ning tekkivad klastrid kujutavad endast sageli pikki ’ahelaid’.

- Täieliku seose meetod ehk kaugeima naabri meetod (*complete linkage method, furthesneigbour method)*

  Siin klastrite vaheline distants suurima kauguse järgi kahe erinevates klastrites asuvate objektide vahel.

  See meetod sobilikum kui objektid moodustavad loomulikke ’kobaraid’. Kui klaster kipub aga olema väljavenitatud või ’keti’ tüüpi, siis see meetod ei sobi.

  Siin kauguse tase, st ühinemise kiirus erinev kuna lähtutakse suurimatest ühinemiskaugustest. Siin ühinemine aeglasem, aga ka pisut selgem.

Tulemus sõltub indiviidide läbivaatamisekorrast. Kui vaadata andmestikku erinevas järjekorras (st kui tegu lähedaste kauguslävedega).

- Keskmise seose meetodid – igal sammul ühendatakse kaks kõige lähemat klastrit.
    - Keskmise gruppidevahelise kauguse meetod (average linkage between groups method) – keskmine kaugus kõikvõimalikest objektipaaridest, üks objekt ühest, teine teisest klastrist Kaugus = keskmise kauguse kõige väiksem kaugus.

    - Keskmine grupisisese kauguse meetod (average linkage within group method) – keskmise vaatlemisel vaadeldakse ka mõlema klastri sees olevaid kaugusi, st arvesse võetakse ka klastrisisene heterogeensus. Kaugus = keskmine kaugus ühendklastri paaridesisese kauguse alusel.

- Ward'i meetod – selle puhul hajuvus klastri sees minimaalseim, seda dispersiooni/hajuvuse mõttes. Klastrite ühendamise kriteeriumiks on see, et ühendatakse need kaks klastrit, mille puhul hajuvuse juurdekasv (dispersiooni mõttes) on vähim. Teatud mõttes see meetod üritab minimiseerida iga kahe (hüpoteetilise) klastri, mis võidakse moodustada igal sammul ruutude summat. Kriteeriumiks olev dispersioon seotud aga keskmisega, so. tsentroididega. Klastrite iseloomustamiseks antud juhul võrreldakse klastrite vahelist dispersiooni ja klastrisisest summaarset dispersiooni. Pole siiski klassikaline dispersiooni kasutamine (F-jaotus), vaid rohkem formaalne suurus. Siin rohkem teatud tõlgendusanaloogia

Mitte-hierarhilised meetodid

- K-keskmiste meetod – kuulub iteratiivsete klasterdamise meetodite hulka.

See meetod erineb oluliselt hierarhilisest meetodist. Oletagem, et on juba ette teada, mitu klastrit peaks meie andmetes olema. Antud juhul on võimalik ’õelda’ arvutile täpselt, et tuleks moodustada näiteks 3 klastrit, mis peaks olema teineteisest võimalikult erinevad. Ehk siis lahendusi on üks, st tuleb ette öelda, mitut klastrit soovitakse (võib tugineda mõnele liigutuspuule mõistliku lahenduse saamiseks).

Üldjoontes annab siis k-keskmiste meetod meile täpselt *k* võimalikult suurima erinevusega klastrit. Oluline on silmas pidada, et parim k number klastreid, mis viib parima eristuse juurde ei ole teada *a priori* vaid see arvutatakse andmetest.

Tehnika koosneb siin üksiksammudest, kusjuures igal sammul parandatakse saadud lahendit. Tulemused tavaliselt klastri koosseisu loeteludena.
Selle meetodi rakendamiseks pole palju ressursse vaja, paljudes programmides nimetatakse seda tehnikat ka kiire klasterdamise meetodiks (quick cluster).

Milles seiseneb esimene samm?

Püüame mõnda elementi klasterdada. Selleks leiame klastri, mille keskpunktile on ta kõige lähemal. Lähima juurde lisamisega muutub ka (selle suhtes) seal tsentroid (x).

Põhiidee – mingil hetkel lastakse kõik elemendid vabaks, st otsitakse, millise klastri tsentroidile on see kõige lähemal ja seal arvutatakse jälle uus tsentroid. Siis lastakse jälle kõik elemendid vabaks jne. ning seda nii kaua kui klastrite keskpunktid enam ei nihku, st elemendid stabiliseeruvad ühe tsentroidi ümber. See klaster koondab endas teatud kontsentratsiooniga elemendid. Teine klaster tõmbab samamoodi oma tsentroidi ümber olevaid objekte.

Arvutuslikult võib sellele meetodile mõelda kui ümberpööratud ANOVA-le (analysis of variance). Programm alustab k juhusliku klastriga ning siis liigutab nende vahel objekte eesmärgiga: 1) minimiseerida klastrite sisest variatiivsust; 2) maksimiseerida variatiivsust klastrite vahel. Teisisõnu, sarnasuse reegleid vastavad maksimaalselt ühe klastri liikmetele ning minimaalsest liikmetele, mis

kuuluvad ülejäänud klastritesse. See on oma idee poolest nagu ümberpööratud ANOVA, mõttes et ANOVA olulisuse testi puhul hinnatakse gruppide vahelist variatiivsust grupisisese variatiivsuse taustal (arvutades olulisuse teste hüpoteesile, et gruppide keskmised on erinevad). K-keskmiste klasterdamise puhul üritab programm liigutada objekte (juhtumeid) gruppidesse ja gruppiest välja (so klastritesse), et saada kõige olulisemad ANOVA tulemused.

Nihke (arvestatava) kriteerium on võimalik ette anda.

Moving centroid (libisev tsentroid) – keskmine/tsentroid arvutatakse kohe iga elemendi lisamisel ümber. Miinuseks see, et see sõltub palju elementide kaasamise järjekorras (nt. erakonda uus liige, kohe vaated muutuvad).

Tulemuste interpreteerimine – tavaliselt k-keskmiste klasteranalüüsi puhul uurime iga klastri keskmist iga dimensiooni puhul, et hinnata kui ’selge/eristuv’ meie klaster on. Ideaalsel juhul saame väga erinevaid keskmisi enamuse kui mitte kõigi analüüsis kasutatud dimensioonidele. F-väärtuste suurus ANOVA-st on teine indikaator näitamaks kui hästi see konkreetne dimensioon eristub klastrite vahel.

Kui palju on aga andmetes klasterid?
Seda ei ole *a priori* teada ning tegelikult ei pruugigi olla konkreetset vastust sellele, millise väärtuse k peaks võtma.

Klasteranalüüsi rakendamise etapid

1) Klasterdatava kogumi määratlus;
2) Tunnuste valik, mille alusel indiviide rühmitatakse;
3) Arvutada indiviidide sarnasuse/erinevuse määr/tase. Milline on sarnasuse definitsioon;
4) Klastrite loomine/leidmine toetudes sarnasuse/erinevuse määrale. Reegli valik klastrite leidmiseks;
5) KA tulemuste tõepärasuse hindamine.

   Kas tulemus on meetodiga leitud või peegeldab see tegelikku olukorda. Kas need langevad kokku. Nagu öeldud, KA-ga saame ettekujutuse struktuurist, kuid ta ei tõesta midagi => vihjete saamine, hüpoteeside püstitamine

LISAKS!!!!

1) Indiviidide arv loodud klastrites peab lõppkokkuvõttes olema piisavalt suur, eriti kui tahetakse tulemusi prognoosida populatsioonileÖ
2) Ilmsed võõrobjektid tuleks klasterdamise eel kõrvaldada (siinkohal nt. standardiseerimine, st muuta andmed võrreldavaks, sest erandlik indiviid viib keskmise liiga suureks, st mõjutab algandmeid ning üldpilt teine kui ilma nende äärmuslike juhtudeta);
3) Missugune kasutatavate koordinaattunnuste arv? Need tunnused peavad sisuliselt sobima liigenduseks, st see on aineteoreetiline otsus.

   Klastrite arv? Proovimine ja sisulised kaalutlused võiksid olla juhiseks, midagi kindlat ei ole.
4) Enamasti luuakse klasterdamisel tunnuste kaalumise võimalus, st tunnustele kaalu omistamise võimalus – mõned tunnused võetakse arvesse 1-st väiksema kaaluga, mõned väärtused õigel kujul = > mõne tunnuse mahasurumine. Näiteks üks tunnus seotud teisega korrelatsiooni mõttes. Need tunnused esindavad mingit sarnast kvaliteeti. Koordinaattuunnustes neid mõlemaid tunnuseid kasutades tuleks klastri struktuur eriti esile (on eriti esilduv).

   Tunnuste vahelisest korrelatsioonist võimalik vabaneda ka peakomponentide meetodiga klasteranalüüsi sees (siin tunnused arvulised, seotud mitte korrelatsiooni alusel)
5) Hinnangute andmine – protsentuaalne tunnuste kaupa analüüs. Klastrite iseloomustamine klasri tunnuste kaudu.