Klusterointimenetelmien validointi

Anna Pienimäki
Anna.Pienimäki@cs.helsinki.fi

Klusterointimenetelmät-seminaari
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C–2002–54, s. 41–51, marraskuu 2002

Tiivistelmä

Klusteroinniksi kutsutaan ohjastamatonta data-alkioiden luokittelua erilaisiin ryhmiin, joita kutsutaan klustereiksi. Klusterointia varten on kehitelty runsas joukko erilaisia algoritmeja, jotka pohjautuvat keskenään hyvin erilaisiin lähtökohtiin. Yhteistä näille algoritmeille on kuitenkin niiden tulosten riippuvaisuus syöteparametreista ja käsiteltävästä datasta.

Jotta erilaisia klusterointimenetelmiä voidaan vertailla ja evaluoida, tarvitaan validointimenetelmiä. Klusteroinnin semanttinen validointi keskittyy arvioimaan klusteroinnin tuloksia eli saatujen klustereiden määrää ja laatua. Optimaalisen klusterointituloksen arviointia varten on kehitetty erilaisia validointimenetelmiä, jotka hyödyntävät niin ennalta annettua optimaalista klusterointiskeemaa kuin myös datan sisäisiä piirteitä. Lisäksi interaktiivisessa validoinnissa on pyritty hyödyntämään ihmisen sisäistä intuitiota datan rakenteesta.

Sekä datan sisäisten piirteiden analysointi että ulkoisen klusterointiskeeman vertailu klusteroinnista saatuun tulokseen vaatii osin raskasta laskentaa. Tämän vuoksi on validoinnin tehostamiseksi pyritty kehittämään erilaisia suhteellisia validointimittoja, joiden avulla voidaan laskea samaa datajoukkoa kohden erilaisille klusterimäärille funktio, jonka käyttäytymistä analysoimalla voidaan päätellä mahdollisimman optimaalinen klusterointiskeema kyseiselle datajoukolle.

Vaikka erilaisia validointimenetelmiä ja -mittoja onkin esitelty runsaasti, ovat ne joko laskennallisesti raskaita tai herkkiä hälylle. Lisäksi validointimittojen tuottamien funktioiden analysoiminen ei aina tuota parasta klusterointitulosta. Siksi optimaalisen klusterointiskeeman löytämistä voidaan edelleen pitää alan avoimena ongelmana.

1 Johdanto

Klusteroinnilla tarkoitetaan keskenään samankaltaisten hahmojen, kuten data-alkioiden tai ominaisuusvektoreiden ohjastamatonta luokittelua ryhmiin, joita kutsutaan klustereiksi [JMF99]. Klusterointia voidaan hyödyntää muun muassa tietämyksen muodostamisen ja tilastollisen data-analyysin tarkoituksiin. Lisäksi klusterointia on hyödynnetty monitieteellisissä sovelluksissa niin laskennallisen biologian kuin musiikkianalyysinkin alalla.

Koska klusterointi on luonteeltaan ohjastamatonta, ei klusterointialgoritmilla ole käytettävissä ennalta määriteltyjä ryhmiä, joihin luokiteltavia alkioita olisi tarkoitus sijoittaa tai tietoutta ominaisuuksista, joiden perusteella luokittelua halutaan suorittaa. Tämän vuoksi klusterointialgoritmien tulokset ovat usein hyvin riippuvaisia sekä niille annetuista parametriarvoista että luokiteltavan datan ominaisuuksista [HBV02a].

Klusterointialgoritmeja voidaan luokitella erilaisten ominaisuuksien, kuten syötedatan, similariteettifunktion kuin algoritmin perusideankin nojalla [HBV01]. Jotta näiden keskenään hyvinkin erilaisten klusterointialgoritmien toimintaa ja tuloksia voitaisiin arvioida ja vertailla, tarvitaan erillisiä validointimenetelmiä. Validointimenetelmät evaluoivat klusterointimenetelmien tuloksia ja pyrkivät vastaamaan kysymyksiin: *montako klusteria datajoukossa on*, *sopiiko löydetty klusterointi datajoukon kuvailuun* ja *voitaisiinko datajoukko klusteroida paremmin* [HBV02b].

Klusteroinnin ohjastamattomasta luonteesta johtuen validointi on hankala tehtävä. Niinpä validointiin on kehitetty useita erilaisia menetelmiä soveltuen erilaisille klusterointimenetelmille ja erilaisia näkökulmia korostaen.

Tässä esitelmässä pohditaan klusteroinnin validoinnin perusongelmaa, esitellään siihen liittyvää käsitteistöä sekä käydään lyhyesti läpi erilaisia klusteroinnin validointimenetelmiä ja niihin liittyviä metriikoita. Esitelmä perustuu pitkälti Halkidi et al. [HBV02a, HBV02b, HBV01] klusteroinnin validointimenetelmien jaotteluun.

Esitelmän toisessa luvussa tarkastellaan lyhyesti hyvän klusterointimenetelmän ominaisuuksia. Luvussa 3 luodaan lyhyt katsaus klusteroinnin semanttisen validoinnin perusongelmaan. Luvussa 4 esitellään erilaisia validointimenetelmiä, joiden avulla erilaisia klusterointimenetelmien tuloksia voidaan evaluoida.

2 Hyvän klusterointimenetelmän ominaisuuksia

Klusterointimenetelmiä voidaan arvioida sekä teknisestä että semanttisesta näkökulmasta. Teknisesti ajatellen hyvän klusterointimenetelmän tulisi olla skaalautuva eli klusterointiin kuluvan ajan tulisi kasvaa vain lineaarisesti tietokannan kokoon nähden. Hyvä klusterointimenetelmä kykenee lisäksi käsittelemään erityyppisiä tietoalkioita joko suoraan ilman muutoksia tai vain hyvin vähäisin muutoksin [ZFL02].

Hyvän klusterointimenetelmän tulisi kyetä tunnistamaan erimuotoisia klustereita, ennen kaikkea moniulotteisessa tilassa. Lisäksi menetelmän tulisi olla mahdollisimman riippumaton aineistoon liittyvästä oheistiedosta ja vaatia mahdollisimman vähän syöteparametreja [ZFL02]. Tästä huolimatta useat klusterointimenetelmät vaativat runsaan määrän syötetietoutta optimaalisen tuloksen saavuttamiseksi.

Käytännön datassa on usein paljon hälyä. Tämän vuoksi hyvän klusterointimenetelmän tulisi kyetä toimimaan hyvinkin hälyisessä ympäristössä. Samoin datan esitysjärjestyksen ei pitäisi vaikuttaa klusterointitulokseen. [ZFL02]

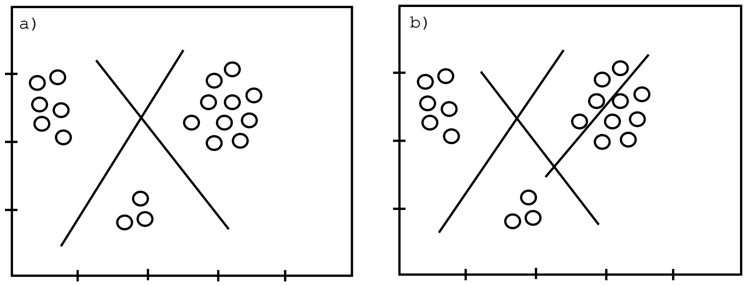
Semanttisesti ajatellen onnistuneen klusterointituloksen tunnusmerkkeinä voidaan pitää klustereiden sisäistä yhteneväisyyttä ja ulkoista erillisyyttä. Tällöin klustereiden lukumäärä olisi mahdollisimman optimaalinen dataan nähden. Yhtenäisen klusterin jäsenten tulisi sijaita mahdollisimman lähellä toisiaan. Yhteneväisyyden mittana käytetään usein varianssia, jonka arvo pyritään minimoimaan [HBV01].

Erillisyydellä tarkoitetaan klustereiden välistä etäisyyttä toisistaan. Erillisyyttä voidaan mitata useamman eri lähestymistavan kautta. Yleisimpiä lähestymistapoja ovat yksittäinen linkki (engl. single linkage), kokonainen linkki (engl. complete linkage) ja klustereiden keskipisteiden vertailu. Yksittäisellä linkillä tarkoitetaan lähestymistapaa, joka mittaa klustereiden etäisyyttä toisistaan niiden toisiaan lähimpien alkioiden välisellä etäisyydellä. Täydellisen linkin tilanteessa tarkastellaan puolestaan klustereiden kauimmaisten alkioiden etäisyyttä. Keskipisteiden tarkastelu mittaa klusterien etäisyyttä niiden keskipisteiden etäisyyksien avulla [HBV01].

3 Klusteroinnin semanttisen validoinnin perusongelma

Klusteroinnin perusongelmana on syötteenä annetun datajoukon kuvailu optimaalisen klusterimäärän avulla. Koska klusterointimenetelmien tulokset ovat useimmiten riippuvaisia sekä syötedatasta että annetuista parametreista, voi sama menetelmä tuottaa erilaisilla parametriarvoilla samasta datasta hyvinkin erilaisen määrän erisisältöisiä klustereita. Vastaavasti jollekin yksittäiselle datajoukolle soveltuvien parametrien käyttö toisen datajoukon klusteroimiseen voi johtaa huomattavan heikkoon klusterointitulokseen [HBV02a].

Vaikka klusterointimenetelmät tuottaisivatkin oikeellisen tuloksen, ei tulos silti välttämättä ole optimaalinen siinä mielessä, että se kuvailisi optimaalisesti annettua datajoukkoa. Tarkastellaan esimerkiksi K-means-algoritmia, joka saa parametrinaan tuloksena haluttujen klustereiden määrän. Tilanteessa, jossa parametri on samansuuruinen datan optimaalisen jakautumisen kanssa, K-means tuottaa yleisesti ottaen varsin hyviä tuloksia. Parametriarvon ollessa joko pienempi tai suurempi kuin optimaalinen datan klusterimäärä, ei saatu tulos kuitenkaan vastaa datan sisäistä jakaumaa (Kuva 3.1).



Kuva 3.1: a) Jaottelu kolmen keskipisteen mukaan. b) Jaottelu neljän keskipisteen mukaan.

Ihmisellä on usein intuitiivinen käsitys hyvästä klusterirakenteesta, jos datajoukko pystytään kuvailemaan tarpeeksi visuaalisesti. Tätä ominaisuutta on pyritty hyödyntämään myös klusteroinnin interaktiivisessa validoinnissa, jossa käyttäjä ohjaa validointiprosessia verifioimalla klusterointialgoritmin tuloksia sen eri vaiheissa. Käytännössä ihminen kykenee kuitenkin hahmottamaan vain hyvin suppeita ulottuvuuksia, minkä vuoksi tarvitaan käyttäjien arvioinneista riippumattomia validointimenetelmiä moniulotteisten datajoukkojen klusteroinnin validointiin [HBV02a].

4 Klusteroinnin validointimenetelmiä

Klusteroinnin validointimenetelmät voidaan karkeasti jakaa kolmeen pääryhmään: validointi ulkoisten, sisäisten ja suhteellisten kriteereiden perusteella. Kolmen tarkasteltavan pääryhmän lisäksi on olemassa myös muita lähestymistapoja klusteroinnin validointiin, mutta ne jätetään tämän esitelmän puitteissa käsittelemättä.

4.1 Nollahypoteesi klusteroinnin validoinnissa

Sekä ulkoisten että sisäisten kriteereiden perusteella tehtävä validointi perustuu tilastolliseen testaukseen. Klusteroinnin validoinnin ollessa käytetään nollahypoteesina H_0 oletusta datajoukon sattumanvaraisesta järjestäytymisestä [HBV02a]. Nollahypoteesi pyritään hylkäämään tai olemaan hylkäämättä tilastollisten testien nojalla. Koska tilastolliset testit ovat luonteeltaan kompleksisia, hyödynnetään sekä ulkoisten että sisäisten kriteereiden perusteella validoitaessa Monte Carlo -tekniikkaa [HBV02a].

Monte Carlon ensimmäisessä vaiheessa generoidaan suuri määrä normaalijakaumaa noudattavia synteettisiä datajoukkoja. Jokaiselle synteettiselle datajoukolle X_i lasketaan käytettävän validointimitan arvo, q_i. Saaduista arvoista q_i laaditaan hajontakaavio, jota voidaan käyttää validointimitan todennäköisyystiheysfunktion approksimaationa.

Validointimittaa vastaavan tiheysfunktion muotoa voidaan tarkastella nollahypoteesin H_0 suhteen. Muoto voi olla kolmenlainen: oikeahäntäinen (engl. right-tailed), vasenhäntäinen (engl. left-tailed) tai kaksihäntäinen (engl. two-tailed). Nollahypoteesin testaamisessa käytetään seuraavia arvostelmia, joissa q on tarkasteltavasta datajoukosta lasketun validointimitan arvo, r on käytettyjen synteettisten datajoukkojen määrä ja ρ on testaukseen valittu kriittinen intervalli:

- Muoto on oikeahäntäinen: H_0 hylätään, jos q on suurempi kuin $(1-\rho)\cdot r$ arvoa kaikista arvoista q_i. Muulloin H_0 jätetään hylkäämättä.
- Muoto on vasenhäntäinen: H_0 hylätään, jos q on pienempi kuin $\rho \cdot r$ arvoa kaikista arvoista q_i. Muulloin H_0 jätetään hylkäämättä.
- Muoto on kaksihäntäinen: H_0 jätetään hylkäämättä, jos q on suurempi kuin $(\rho/2)\cdot r$ arvoa kaikista arvoista q_i ja pienempi kuin $(1-\rho/2)\cdot r$ arvoa kaikista arvoista q_i.

4.2 Ulkoisten kriteereiden perusteella validointi

Ulkoisten kriteereiden avulla tehtävä validointi [HBV02a, HBV01] voidaan jakaa kahteen perustekniikkaan: klusterointimenetelmällä tuotetun klusterointirakenteen C alkioittainen vertailu datasta intuitiivisesti muodostettuun ja validointimenetelmälle syötettyyn jaotteluun P sekä datajoukon X pisteiden etäisyysmatriisin E vertailu jaotteluun P.

4.2.1 Klusterointiskeeman C vertailu annettuun jaotteluun P

Sovellettaessa klusterointituloksen $C = \{C_1 ... C_m\}$ vertailua jaotteluun $P = \{P_1 ... P_s\}$ jaotellaan kukin syötedatan pistepari (x_v, x_u) seuraaviin ryhmiin:

- *SS*: Jos molemmat pisteet kuuluvat samaan klusteriin sekä C:ssä että P:ssä
- *SD*: Jos molemmat pisteet kuuluvat samaan klusteriin C:ssä mutta eri klustereihin P:ssä
- *DS*: Jos molemmat pisteet kuuluvat eri klustereihin C:ssä mutta samaan klusteriin P:ssä
- *DD*: Jos molemmat pisteet kuuluvat eri klustereihin sekä C:ssä että P:ssä

Oletetaan että a, b, c ja d ovat pisteparien lukumäärät joukoissa SS, SD, DS, DD. Lasketaan arvo $M = a + b + c + d$, joka on pisteparien maksimilukumäärä. Näiden avulla voidaan määritellä esimerkiksi seuraavat yksinkertaiset validointimitat:

- Rand: $R = (a + d)/M$,
- Jaccard: $J = a/(a + b + c)$

Sekä Randin että Jaccardin mitta saavat arvoja väliltä 0..1 arvon 1 tarkoittaessa täysin yhteneviä klusterointiskeemoja.

Klusterointiskeemoja C ja P voidaan tarkastella myös matriisimuotoisina M_{ij}, jossa matriisialkio m_{ij} on kahden datapisteen suhde matriisin kuvaamassa klusterointiskeemassa. Nyt voidaan käyttää matriisien yhteneväisyyden laskentaan Hubertin Γ -mittaa, joka voidaan myös normalisoida haluttaessa.

- Hubertin Γ:
$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{ij} p_{ij}$$

Esimerkki 1 Tarkastellaan nollahypoteesin testausta edellä esiteltyjen mittojen avulla. Oletetaan, että alkuperäisenä klusteroitavana datajoukko X koostuu sadasta kolmiulotteisesta datapisteestä. Oletetaan lisäksi, että intuitiivisesti ryhmiteltynä klusterointiskeemaksi P datapisteet muodostavat järjestyksessä neljä klusteria, joista jokainen sisältää 25 pistettä. Nyt klusteri P_a sisältää ensimmäiset 25 pistettä, toiset 25 kuuluvat klusteriin P_b, kolmannet klusteriin P_c ja loput klusteriin P_d.

Tarkastellaan esimerkiksi K:n keskipisteen klusterointimenetelmän validointia [Hak01, s. 349]. Suoritetaan klusterointi antaen parametrina k algoritmille halutuksi klustereiden määräksi intuitiivista klusterointiskeemaa vastaava $k = 4$. Nyt voidaan laskea saadulle tulosskeemalle C ja annetulle jaottelulle P Randin, Jaccardin ja Hubertin Γ -mitat.

Jotta nollahypoteesin suhteen voitaisiin tehdä jotain päätelmiä, generoidaan Monte Carlo -tekniikalla approksimaatiot mittojen todennäköisyystiheysfunktioista seuraavasti. Generoidaan 100 datajoukkoa $X_i, i = 1, ..., 100$, joista jokainen sisältää X:ää mallintaen sata kolmiulotteista datapistettä. Jokaista synteettistä joukkoa X_i kohden suoritetaan samat operaatiot kuin alkuperäiselle datajoukolle eli jaetaan joukko vastaavasti neljään osaan skeemaksi $P_i = \{P_{i_a}, ..., P_{i_d}\}$ ja suoritetaan K:n keskipisteen klusterointi parametrilla $k = 4$, minkä tuloksena on skeema C_i. Kullekin C_i ja P_i, $i = 1, ..., 100$, lasketaan vastaavat mitat kuin alkuperäisen datan klusterointiskeemoille.

Valikoidaan seuraavaksi kriittinen intervalli ρ, olkoon esimerkiksi $\rho = 0.05$. Koska Rand, Jaccard ja Hubertin Γ tuottavat oikeahäntäisen tiheysfunktion muodon, tarkastellaan, ovatko alkuperäisen datajoukon arvot R, J ja Γ suurempia kuin $(1-0.05)100 = 95$ arvoa arvoista R_i, J_i ja Γ_i. Jos arvot R, J ja Γ ovat suurempia kuin yli 95 kpl arvoista R_i, J_i ja Γ_i, nollahypoteesi hylätään eli alkuperäisen datajoukon katsotaan sisältävän jonkinlaisen satunnaisesta poikkeavan rakenteen. Muussa tapauksessa nollahypoteesi jätetään hylkäämättä ja datan katsotaan olevan satunnaisesti rakenteinen.

4.2.2 Etäisyysmatriisin E vertailu annettuun jaotteluun P

Klusterointiskeeman C sijaan voidaan tarkastella haluttaessa datajoukon X pisteistä muodostettua etäisyysmatriisia E. Etäisyysmatriisia E voidaan vertailla klusterointiskeeman P kanssa seuraavasti. Tarkastellaan jaottelua P relaationa $g : X \to \{1...n_c\}$, jossa n_c on klustereiden lukumäärä. Määritellään skeemasta P matriisi Y,

$$Y(i,j) = \begin{cases} 1 & , \text{ jos } g(x_i) \neq g(x_j), i, j = 1..N \\ 0 & , \text{ muuten} \end{cases}$$

Etäisyysmatriisin E ja jaottelusta saadun matriisin Y vertailuun voidaan käyttää esimerkiksi edellä esiteltyä Hubertin Γ-arvoa tai sen normalisoitua muotoa.

Esimerkki 2 Monte Carlo -tekniikkaa sovelletaan matriisilähestymistapaan seuraavasti. Oletetaan jälleen, että alkuperäinen datajoukko X sisältää 100 kolmiulotteista datapistettä, jotka jakautuvat esimerkin 1 tavoin neljäksi klusteriksi $P = \{P_a, ..., P_d\}$. Datajoukolle X muodostetaan sen etäisyysmatriisi E ja jaottelua P vastaava matriisi Y. Näille lasketaan esimerkiksi Hubertin Γ -arvo, Γ. Generoidaan jälleen sata synteettistä datajoukkoa X_i, joista jokaiselle laaditaan sekä etäisyysmatriisi E_i että jaottelua P_i vastaava matriisi Y_i. Näille kullekin lasketaan Γ_i. Tämän jälkeen voidaan tarkastella saatua todennäköisyystiheysfunktion approksimaatiota ja joko hyljätä tai jättää hylkäämättä nollahypoteesi sen nojalla, ovatko synteettiset Γ-arvot alkuperäistä pienempiä vai suurempia asetetun kriittisen intervallin suhteen.

4.3 Sisäisten kriteereiden perusteella validointi

Sisäisten kriteereiden perusteella validoitaessa [HBV02a, HBV01] kaikki validointiin tarvittavat tiedot eristetään datajoukosta itsestään. Sisäisin kriteerein validointi voidaan jakaa kahteen eri tapaukseen: hierarkkisten klusterointialgoritmien validointiin ja ei-hierarkkisten klusterointialgoritmien validointiin.

4.3.1 Hierarkkisten klusterointialgoritmien validointi

Hierarkkisen klusterointialgoritmin ollessa kyseessä voidaan laskea matriisi P_c, jonka kukin alkio ilmaisee sitä, millä etäisyystasolla pisteet x_i ja x_j ovat löytyessään samasta klusterista ensimmäisen kerran klusterointialgoritmin suorituksen aikana. Klusteroitavaa datajoukkoa vastaavan etäisyysmatriisi E:n ja P_c:n vertailu voidaan suorittaa laskemalla mitta CPCC (engl. Cophenetic Correlation Coefficient), joka määritellään seuraavasti:

$$CPCC = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} e_{ij}c_{ij} - \mu_E\mu_C}{\sqrt{([(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} e_{ij}^2 - \mu_E^2][(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_C^2]}},$$

jossa $M = N * (N-1)/2$ on datapisteparien lukumäärä, kun N on kaikkien datapisteiden lukumäärä klusteroitavassa datajoukossa X ja e_{ij} ja c_{ij} ovat matriiseiden E ja P_c alkioita.

$$\mu_E = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E(i,j), \mu_C = (1/M)\sum_{i=1}^{N}\sum_{j=i+1}^{N} P_c(i,j)$$

.

CPCC:n arvo vaihtelee välillä $-1 \leq CPCC \leq 1$. Mitä suuremman arvon CPCC saa, sen samankaltaisempia ovat matriisit E ja P_c.

Myös hierarkkisia klusterointialgoritmeja sisäisin kriteerein validoitaessa voidaan käyttää edellä esiteltyä Monte Carlo -tekniikkaa nollahypoteesin testaamiseen. Tällöin matriisit E_i ja P_{c_i} sekä niistä saatu $CPCC_i$-arvo lasketaan jokaista synteettistä datajoukkoa kohden. Hypoteesin hylkäämisen suhteen menetellään kuten ulkoisin kriteerein validoitaessa.

4.3.2 Ei-hierarkkisten klusterointialgoritmien validointi

Validoitaessa ei-hierarkkista klusterointialgoritmia voidaan vertailla klusterointitulosta C datajoukkoa kuvailevaan etäisyysmatriisiin E. Klusterointiskeemalle C lasketaan nyt vastaavanlainen matriisi Y kuin ulkoisin kriteerein validoitaessa jaottelulle P. Tämän avulla voidaan laskea tarvittu Hubertin Γ-arvo. Monte Carlo -menetelmää käytetään kuten ulkoisten kriteereiden tapauksessa.

4.4 Suhteellisten kriteereiden perusteella validointi

Koska sekä ulkoisin että sisäisin kriteerein validointi on laskennallisesti hyvin raskasta ja tarjoaa lähinnä vastauksen kysymykseen, onko saatu klusterimäärä soveltuva dataan ilman määrittelyä paremmasta klusterimäärästä, on kehitetty myös toisenlaisia, suhteellisiin kriteereihin perustuvia validointimenetelmiä [HBV02b, HBV01].

Suhteellisten kriteereiden perusteella tehtävän validoinnin perusideana on klusteroida sama data useita kertoja samalla algoritmilla vaihtaen algoritmin saamia parametriarvoja. Saatujen klusterointitulosten perusteella arvioidaan, mikä klustereiden määrä on todennäköisimmin optimaalinen datajoukon kuvailuun.

Suhteellisten kriteereiden perusteella tehtävä validointi jakautuu kahteen perustapaukseen: haluttu klustereiden määrä joko on tai ei ole annettu parametrina algoritmille. Jos haluttua klustereiden määrää ei ole annettu parametrina, algoritmia suoritetaan hyvin suurella joukolla erilaisia parametreja ja valitaan laajin sellainen parametriarvoalue, joissa klustereiden määrä pysyy vakiona. Muiden parametrien suositeltavat arvot valitaan asetettujen arvoalueiden keskikohdista.

Jos haluttu klustereiden määrä on annettu parametrina, käytetään parhaimman klusteroinnin arvioimisessa validointimittaa q. Validointi tapahtuu nyt seuraavasti.

- Klusterointialgoritmia suoritetaan valikoimalla halutuksi klusterimääräksi n_c vuorollaan kaikki arvot käyttäjän antamien minimi- ja maksimiarvojen väliltä.
- Jokaista annettua klusterimääräarvoa $n_c$kohden algoritmi suoritetaan r kertaa vaihdellen algoritmin vaatimien muiden parametrien arvoja.
- Parhaimmat mitan q saamat arvot kutakin klusterimäärää n_c kohden piirretään n_c:n funktiota kuvaavaksi kaavioksi.

Kaavion tulkinnassa tarkastellaan klusterimäärää kuvaavan funktion käyttäytymistä. Jos klusterimäärän noustessa funktion arvot pysyttelevät samalla tasolla, etsitään funktion maksimi- tai minimikohta. Jos puolestaan funktio joko nousee tai laskee klusterimäärän noustessa, etsitään sellaisia klusterimääräarvoja, joiden kohdalla on merkittävä lokaali vaihtelu funktiossa. Lokaalin vaihtelun puute voi tällöin merkitä klusteroitavan datan rakenteen satunnaisuutta.

Tarkastellaan seuraavaksi lähemmin muutamaa validointimittaa sekä erotteleville että limittäisille klusterointimenetelmille.

4.4.1 Erotteleva klusterointi

Jos validoitava menetelmä klusteroi datapisteet siten, että kukin piste kuuluu yhteen ja vain yhteen klusteriin, puhutaan erottelevasta klusteroinnista (engl. crisp clustering). Erottelevan klusteroinnin validointiin suhteellisin kriteerein on kehitetty runsas joukko erilaisia arviointimittoja variaatioineen. Tarkastellaan seuraavassa niistä muunneltua Hubertin Γ-arvoa, Dunn-mittaa, Davies-Bouldin mittaa ja SD-validointimittaa.

Muunneltu Hubertin Γ-arvo voidaan määritellä yhtälöllä

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_{ij} \cdot q_{ij},$$

jossa $M = N(N-1)/2$, E kuvaa datajoukon etäisyysmatriisia ja Q on $N \times N$ matriisi, jonka kukin alkio q_{ij} kuvastaa niiden klustereiden etäisyyttä, joihin alkiot x_i ja x_j kuuluvat. Klustereiden etäisyys lasketaan pisteiden (v_{ci}, v_{cj}) etäisyyksinä, näiden ollessa joko klustereiden keskipisteitä, niiden kauimpia tai lähimpiä pisteitä.

Myös muunnellusta Hubertin Γ-arvosta voidaan käyttää normalisoitua muotoa. Hubertin Γ-arvo on korkea silloin, kun data sisältää kiinteitä klustereita, mikä näkyy validointifunktiossa tapahtuvana merkittävänä lokaalina vaihteluna optimaalisen klusterointimäärän kohdalla.

Dunn-mitta pyrkii löytämään kiinteitä ja hyvin eroteltuja klustereita. Dunn-mitta määritellään seuraavasti:

$$D_{n_c} = min_{i=1,...,n_c} \left\{ min_{j=i+1,...,n_c} \left(\frac{d(c_i, c_j)}{max_{k=1,...,n_c} diam(c_k)} \right) \right\},$$

jossa $d(c_i, c_j) = min_{x \in c_i, y \in c_j} d(x, y)$ on etäisyysfunktio klustereille c_i ja c_j ja $diam(C) = max_{x,y \in C} d(x, y)$ on klusterin C halkaisija. Koska kiinteiden ja hyvin eroteltujen klustereiden halkaisijat ovat pieniä ja klustereiden etäisyydet suuria, ilmaisee suuri Dunn-arvo hyvää klusterointitulosta. Dunn-mittaa käytettäessä etsitäänkin saadusta validointikaaviosta se klusterointimäärä, jolle mitta saa korkeimman arvonsa. Koska Dunn-mitta on laskennallisesti raskas ja suhteellisen herkkä hälylle, on sille laadittu useita variaatioita, jotka jätetään kuitenkin tämän esitelmän puitteissa käsittelemättä.

Davies-Bouldin -mitta kertoo klustereiden keskimääräisen similariteetin, joten sitä käytettäessä parhain mahdollinen klusterointi on se, jolla mitta saa mahdollisimman pienen arvon. Kahden klusterin C_i ja C_j välisen similariteettifunktion R_{ij} tulee saada ei-negatiivisia arvoja ja olla symmetrinen. Yksinkertainen malli tällaiselle funktiolle on esimerkiksi: $R_{ij} = (s_i + s_j)/d_{ij}$, jossa s_i, s_j ovat klustereiden C_i, C_j keskihajonnat ja d_{ij} on klustereiden välinen etäisyys.

Tällöin Davies-Bouldin -mitta voidaan määritellä seuraavasti:

$$DB_{n_c} = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i,$$

jossa

$$R_i = max_{i=1,...,n_c, i \neq j} R_{ij}, i = 1, ..., n_c$$

ja n_c on klustereiden määrä.

Koska Davies-Bouldin saa pienimmän arvonsa, kun klusterit ovat mahdollisimman erilaisia keskenään, etsitään mittaa käytettäessä validointikaaviosta se klusterimäärä, jolle mitta saa pienimmän arvonsa. Myös Davies-Bouldin -mitalle on kehitetty erilaisia variantteja mutta ne rajataan Dunn-mitan varianttien tavoin tarkastelun ulkopuolelle.

SD-mitta voidaan määritellä käsitteiden *klustereiden keskimääräinen hajonta* ja *klustereiden välinen totaalinen erotus* avulla. Klustereiden keskimääräinen hajonta määritellään seuraavasti:

$$Scat(n_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \|\sigma(v_i)\| / \|\sigma(X)\|,$$

jossa $\sigma(v_i), i = i, ..., n_c$ on kunkin klusterin ja $\sigma(X)$ datajoukon keskihajonta. $Scat(n_c)$ mittaa klustereiden sisäistä kompaktiutta antaen pienen arvon klusterointiskeemalle, jonka klusterit ovat kompakteja ja suuren skeemalle, jonka klustereiden sisäinen keskihajonta on suuri. Klustereiden välinen totaalinen erotus puolestaan määritellään:

$$Dis(n_c) = \frac{D_{max}}{D_{min}} \sum_{k=1}^{n_c} \left(\sum_{z=1}^{n_c} \|v_k - v_z\| \right)^{-1},$$

jossa D_{max} on klustereiden keskipisteiden maksimietäisyys ja D_{min} minimietäisyys. $Dis(n_c)$ mittaa klustereiden etäisyyttä toisiinsa ja tyypillisesti klusterimäärän kasvaessa myös sen arvo kasvaa.

Näiden perusteella voidaan määritellä SD-mitta seuraavasti: $SD(n_c) = a \cdot Scat(n_c) + Dis(n_c)$, jossa $a = Dis(c_{max})$ ja c_{max} on klustereiden maksimimäärä. Termiä a käytetään tasapainottamaan yhtälön jäsenten suhdetta toisiinsa.

Esimerkki 3 Oletetaan, että validoidaan klusterointialgoritmia, joka saa halutun klusterimäärän parametrinaan (esimerkiksi K:n keskipisteen algoritmia). Tarkastellaan jälleen sadasta datapisteestä koostuvaa datajoukkoa X, joka luontevasti jakautuisi neljään klusteriin. Oletetaan lisäksi, että validointia varten on määritelty klusterimäärälle minimi- ja maksimiarvot $n_{c_{min}} = 2, n_{c_{max}} = 5$.

Käytettäessa K:n keskipisteen algoritmia kutakin parametriarvoa $n_c = 2, ..., 5$ kohden klusterointialgoritmi ajetaan kertaalleen, sillä K:n keskipisteen algoritmi ei saa muita parametreja kuin datajoukon X ja klusterimäärän [Hak01, s. 349]. Jos algoritmi saisi muita parametreja, suoritettaisiin algoritmi r kertaa, r:n ollessa esimerkiksi $r = a_i \times a_j \times a_k$, jossa a_i, a_j, a_k ovat muiden parametrien arvojoukkojen kokoja.

Jokaselle klusterimäärälle $n_c = 2, ..., 5$ lasketaan r klusterointiskeemaa $C_i, i = 1, ..., r$, joista puolestaan jokaiselle lasketaan haluttu validointimitta, olkoon esimerkiksi $Dunn_i$. Kutakin n_c kohden valikoidaan paras tuotettu arvo, Dunn-mitan tapauksessa suurin löydetty arvo. Näistä tuotetaan n_c:n funktio, jonka maksimikohdasta löydetään paras mahdollinen klusterimäärä sekä siihen liittyvät suositusparametriarvot.

4.4.2 Limittäinen klusterointi

Klusterointia, joka sallii klusteroitavien data-alkioiden kuulumisen yhtäaikaisesti useampaan klusteriin, kutsutaan limittäiseksi klusteroinniksi (engl. fuzzy clustering). Limittäiset klusterit salliville menetelmille kehitettyjä validointimittoja voidaan käyttää kuten yllä mainittuja mittoja luoden klustereiden määrälle funktio ja analysoiden tämän maksimeja, minimejä tai lokaaleja vaihteluita. Limittäiset klusterit sallivien menetelmien kohdalla päämääränä on etsiä klusterointi, jossa suurimmalla osalla data-alkioista on korkea jäsenyysarvo yhdessä klusterissa. Limittäiset klusterit sallivien menetelmien tuottamat klusteroinnit voidaan esittää matriisimuodossa matriisilla $U = [u_{ij}]$, missä u_{ij} on alkion x_i jäsenyys j:nnessä klusterissa.

Limittäiset klusterit sallivien menetelmien validointimitat voidaan jakaa kahteen ryhmään: mittoihin, jotka käyttävät ainoastaan matriisia U ja mittoihin, jotka käyttävät sekä matriisia U että datajoukkoa itseään. Pelkkää U-matriisia käyttäviä mittoja on kaksi: PC (engl. partition coefficient) ja PE (engl. partition entropy coefficient). PC määritellään seuraavasti

$$PC = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_c} u_{ij}^2$$

ja sen arvo vaihtelee välillä $[1/n_c, 1]$ n_c:n ollessa klustereiden määrä ja u_{ij}:n matriisin U alkio (i, j). Mitä suuremman arvon mitta saa, sen vähemmän klusterit limittyvät. Toisin sanoen, mitä pienempi arvo PC-mitalla on, sen vähemmän datajoukosta löytyy klusteroituvaa rakennetta tai klusteroinnin tuottanut algoritmi ei ole kyennyt sitä löytämään.

PE määritellään puolestaan

$$PE = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_c} u_{ij} \cdot log_a(u_{ij})$$

ja sen arvo vaihtelee välillä $[0, log_a n_c]$ pienten arvojen tarkoittaessa vahvaa ja suurten heikkoa klusteroitumista.

Kirjallisuus esittelee koko joukon mittoja, jotka hyödyntävät sekä U-matriisia, että datajoukon piirteitä. Tässä esitelmässä niistä esitellään esimerkin vuoksi vain Xie-Beni -mitta, joka määritellään seuraavasti:

$$XB = \pi / N \cdot dmin,$$

jossa N on data-alkioiden määrä ja dmin on klustereiden pienin etäisyys, $dmin = min\|v_i - v_j\|$, jossa v_i, v_j ovat klustereiden keskipisteitä. π puolestaan määritellään skeeman klustereiden kompaktiudeksi. Koska hyvässä klusteroinnissa π saa pieniä ja $N \cdot dmin$ suuria arvoja eli klusterit ovat kompakteja ja etäällä toisistaan, tulee XB-arvon olla mahdollisimman pieni.

5 Yhteenveto

Erilaisten klusterointimenetelmien määrä on kasvanut runsaasti tietämyksen muodostamisen alan kasvun myötä. Koska klusterointimenetelmät ovat keskenään hyvinkin erilaisia ja useimmat lisäksi hyvin riippuvaisia saamistaan syöteparametreista ja syötedatan ominaisuuksista, tarvitaan menetelmiä näiden evaluointiin ja validointiin.

Validointia voidaan tehdä sekä teknisestä että semanttisesta näkökulmasta. Näistä semanttinen näkökulma kiteytyy pitkälti klustereiden optimaalisen lukumäärän löytämiseen ja datajoukon mahdollisimman oikeelliseen kuvaamiseen. Semanttisesta näkökulmasta tehtävään validointiin kehitetyt menetelmät voidaan jaotella ryhmiin sen perusteella, käyttävätkö ne ulkoista määrittelyä optimaaliselle datan klusteroinnille, datajoukosta eroteltavia piirteitä vai pyritäänkö optimaalinen klusterointi löytämään eri parametriarvoja testaamalla.

Sekä ulkoisin että sisäisin kriteerein tehtävä validointi on laskennallisesti raskasta. Siksi on pyritty löytämään erilaisia suhteellisia validointimittoja, joiden avulla voidaan määritellä klusterimäärälle funktio, jonka maksimikohtia ja lokaalisia vaihteluita tarkastelemalla pyritään löytämään optimaalinen klusterointiskeema. Erilaisten validointimittojen avulla pystytään evaluoimaan niin limittäisiä kuin erillisiäkin klustereita tuottavia menetelmiä.

Vaikka kirjallisuus esitteleekin runsaan joukon erilaisia validointimenetelmiä, ovat ne tyypillisesti laskennallisesti raskaita ja aikaavieviä. Siksi optimaalisen klusterointiskeeman löytämistä voidaan edelleen pitää eräänä alan avoimista kysymyksistä.

Viitteet

[HBV02a] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering Validity Checking Methods: Part I. *SIGMOD Record*, June 2002.

[HBV02b] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering Validity Checking Methods: Part II. *SIGMOD Record*, September 2002.

[HBV01] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techiques. *Journal of Intelligent Information Systems*, 17(2-3), pages 107–145, 2001.

[Hak01] J. Han and M. Kamber. *Data Mining – Concepts and Techniques*. Academic Press, 2001.

[JMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3), pages 264–323, September 1999.

[ZFL02] O. R. Zaiane, A. Foss, C.-H. Lee, and W. Wang. On Data Clustering Analysis: Scalability, Constraints and Validation. In *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02)*, Taipei, Taiwan, May 2002.