Klusterointimenetelmät

Marko Tuononen
Joensuun yliopisto, Tietojenkäsittelytiede

Laudaturseminaariesityksen kirjallinen tukimateriaali

25. helmikuuta 2005

Tiivistelmä

Klusterointi on menetelmä saada tietoa aineiston rakenteesta. Klusterointi on tärkeä työkalu muun muassa monilla tietojenkäsittelytieteen osa-alueilla sekä useiden empiiristen tieteiden parissa. Tässä työssä annetaan yleisesittely klusteroinnista sekä muutamia esimerkkejä mahdollisista sovelluksista. Lisäksi esitellään luokittelu klusterointimenetelmille ja luodaan yleiskatsaus klusterointimenetelmiin.

ACM-luokat (ACM Computing Classification System, 1998 version): I.5.3

Avainsanat: klusterointi, klusterointimenetelmät, klusterointimenetelmien luokitus

1 Johdanto

Annettujen aineellisten tai abstraktien alkioiden jaottelemista ryhmiin samankaltaisuuden perusteella kutsutaan klusteroinniksi (Han & Kamber, 2001). Asioiden ja ilmiöiden jaottelu ryhmiin on ihmisille erittäin luonteenomaista toimintaa. Ihminen oppii jo aikaisessa lapsuudessaan erottamaan kissan koirasta sekä löytämään eron kasvi- ja eläinkunnan väliltä jatkaen läpi elämänsä tämän alitajuisen jaottelun tarkentamista.

Klusteroinnilla on pitkä historia useissa empiirisissä tieteissä, kuten esimerkiksi biologiassa, psykiatriassa, psykologiassa, lääketieteessä, arkeologiassa, geologiassa, sosiologiassa ja markkinoinnissa, joissa klusterointia usein käytetään datan analysointiin (Jain *& al.*, 1999). Klusteroinnilla on useita sovelluksia myös tietojenkäsittelytieteen alalla, sillä klusterointi on tärkeä työkalu muun muassa *hahmontunnistuksessa* (pattern recognition), *data analyysissä* (data analysis), *kuvankäsittelyssä* (image processing) sekä *tiedonlouhinnassa* (data mining) (Han & Kamber, 2001). Seuraavassa on esitelty muutamia klusteroinnin sovelluksia biologian, psykiatrian sekä markkinoinnin aloilta:

- Biologiassa klusterointia on voitu käyttää kasvi- ja eläinluokittelujen muodostamisessa sekä populaatioiden sisäisen rakenteen tutkimisessa (Han & Kamber, 2001). Esimerkiksi Fränti ja muut (2000) ovat käyttäneet klusterointia bakteerien luokitteluun. Lisäksi geenien *ekspressiodataa* klusteroimalla on voitu löytää toiminnaltaan samankaltaiset geenit (Seal *& al.*, 2005).
- Psykiatriassa on klusterointia käyttäen muun muassa pyritty luokittelemaan itsemurhaan taipuvaisia mielenterveyspotilaita. Koska mielenterveysongelmat ovat monimutkaisempia kuin muut terveyteen liittyvät ongelmat, on psykiatriassa yleisemminkin osoitettu mielenkiintoa klusteroinnin soveltamiseksi nykyisten ongelmaluokkien tarkentamiseksi tai jopa uudelleen määrittelemiseksi (Everitt, 1993).
- Markkinoinnissa voidaan kuluttajien ostokäyttäytymistä klusteroimalla löytää erilaisia kuluttajaryhmiä. Tätä tietoa voidaan hyödyntää suunniteltaessa täsmämainontaa (Han & Kamber, 2001). Markkinointitutkimuksissa halutaan yleensä kerätä tietoa erilaisilta kuluttajaryhmiltä. Klusteroimalla voidaan jakaa

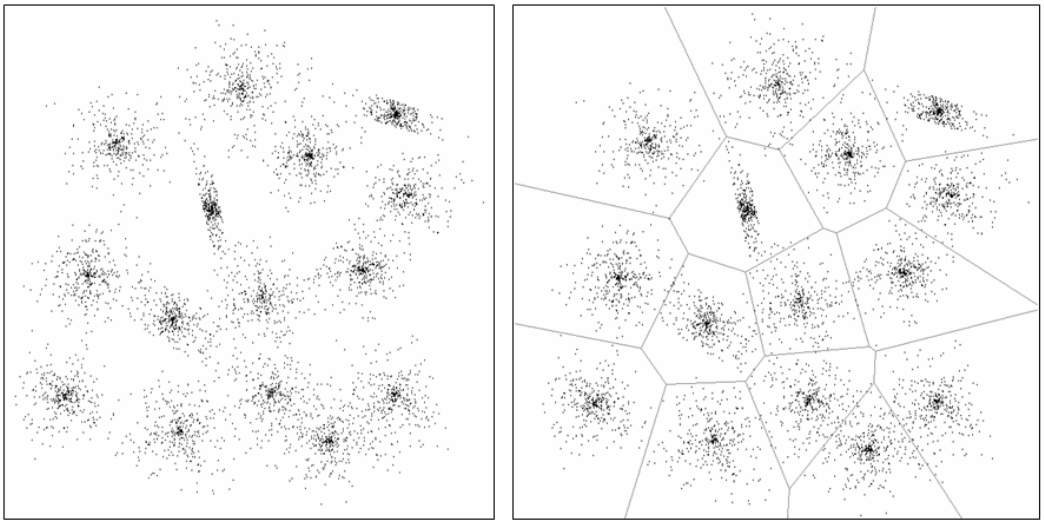
kaupungit tai kaupunginosat samankaltaisiin ryhmiin, jolloin voidaan paremmin varmistua kyselyiden kohdistumisesta tasapuolisesti kaikkiin kuluttajaryhmiin (Everitt, 1993).

2 Klusterointi

Olkoon annettu joukko aineellisia tai abstrakteja alkioita. Kutsumme jatkossa tätä annettujen alkioiden joukkoa aineistoksi. *Klusteroinnissa* (clustering) pyritään aineistosta löytämään ryhmiä ja jakamaan alkiot näihin ryhmiin siten, että alkiot kussakin ryhmässä ovat keskenään mahdollisimman samanlaisia mutta alkiot eri ryhmissä ovat keskenään mahdollisimman erilaisia (Kaufman & Rousseeuw, 1990; Everitt, 1993; Han & Kamber, 2001). Muodostettuja ryhmiä kutsutaan klustereiksi. *Klusteri* (cluster) määritellään aineiston epätyhjänä osajoukkona. Lisäksi vaaditaan, ettei millään kahdella klusterilla ole yhteisiä alkioita. Poikkeuksena tästä on *sumeaan logiikkaan* (fuzzy logic) perustuva klusterointi, jossa kukin alkio voi kuulua useampaan kuin yhteen klusteriin (Jain *& al.*, 1999). Jätämme kuitenkin tässä työssä huomiotta sumeaan logiikkaan perustuvat klusterointimenetelmät.

Klusteroitava *aineisto* voi olla esimerkiksi joukko kuvia, jotka halutaan järjestää sisältönsä mukaan (Chen *& al.*, 2000) tai geenien ekspressiodataa, jota klusteroimalla halutaan löytää toiminnaltaan samankaltaiset geenit (Seal *& al.*, 2005). Jotta klusterointi olisi mielekästä, oletamme, että klusteroitavalla aineistolla on jonkinlainen rakenne. Voidaksemme klusteroida tarvitsemme lisäksi keinon mitata aineiston alkioiden erilaisuutta tai samankaltaisuutta. Jatkossa oletamme aineiston olevan joukko tason pisteitä ja aineiston alkioiden erilaisuusmittana käytettävän *euklidista etäisyyttä*.

Kuvassa 1 on esitetty esimerkki klusteroinnista, jossa klusteroitavana aineistona on 5000 tason pistettä. Muodostetut 15 klusteria on osoitettu rajaamalla syntyneet klusterit viivoilla. Rajaavat viivat on saatu piirtämällä *Voronoin diagrammi* klustereiden keskipisteiden suhteen (Aurenhammer, 1991).



Kuva 1: Eräs aineisto ja sen klusterointi.

Optimaalisen klusteroinnin muodostaminen on kombinatorisessa muodossaan *NP-täydellinen* ongelma (Garey *& al.*, 1982), joten tehokkaimmatkin optimaalisen ratkaisun tuottavat menetelmät, kuten esimerkiksi Fräntin ja Virmajoen (2004) esittämä menetelmä, ovat aikavaativuudeltaan eksponentiaalisia. Näin ollen on käytännössä käytettävä klusterointimenetelmiä, jotka eivät tuota optimaalista ratkaisua. Todennäköisesti johtuen juuri tästä optimaalisen klusteroinnin muodostamisen vaikeudesta sekä klusteroinnin sovelluskentän laajuudesta, on olemassa hyvin monenlaisia lähestymistapoja ei-optimaalisen klusteroinnin muodostamiseksi.

Klusterointimenetelmien luokittelemiseksi ei kirjallisuudesta löydy mitään yhtenäistä linjaa. Esimerkiksi Han ja Kamber (2001), Jain ja muut (1999), Ma ja Chow (2004) sekä Virmajoki (2004) esittävät kukin erilaisen luokittelun klusterointimenetelmille. Tässä työssä esitettävässä klusterointimenetelmien luokittelussa on käytetty soveltuvin osin edellä mainittuja luokitteluja.

Kirjallisuudessa on esitetty lukemattomia klusterointimenetelmiä. Klusterointimenetelmän valinnassa on otettava huomioon klusteroitavan aineiston tyyppi sekä klusteroinnin käyttötarkoitus (Han & Kamber, 2001). Mikäli klusterointia käytetään *tutkivana menetelmänä* (exploratory tool), voidaan aineistoa klusteroida useilla eri menetelmillä, jotta nähdään, mitä aineistosta paljastuu (Han & Kamber, 2001; Jain *& al.*, 1999). Klusterointimenetelmät voidaan karkeasti jakaa hierarkkisiin ja osittaviin menetelmiin.

3 Hierarkkiset menetelmät

Hierarkkiset menetelmät tuottavat aineistolle useita klusterointeja, jotka muodostavat puumaisen hierarkian. Muodostettu hierarkia voidaan esittää *dendrogrammina* (dendrogram). Hierarkkiset menetelmät jaetaan edelleen jakaviin ja yhdistäviin menetelmiin. *Jakavissa menetelmissä* (divisive) aineiston kaikki alkiot ovat aluksi samassa klusterissa. Tämän jälkeen klustereita jaetaan kahtia kunnes saavutetaan haluttu määrä klustereita tai kunnes kukin alkio on omana klusterinaan. *Yhdistävissä menetelmissä* (agglomerative) kukin aineiston alkio on aluksi omana klusterinaan. Tämän jälkeen klustereita yhdistellään kunnes saavutetaan haluttu määrä klustereita tai kunnes kaikki alkiot ovat samassa klusterissa. (Jain *& al.*, 1999; Kaufmann & Rousseeuw, 1990; Everitt, 1993)

Jakavien menetelmien tapauksessa on ratkaistava, miten valitaan jaettava klusteri sekä miten valittu klusteri jaetaan. Jaettava klusteri voidaan valita esimerkiksi ryhmän koon (Kaufmann & Rousseeuw, 1990) tai *hajonnan* (variance) (Wu & Zhang, 1991; Fränti *& al.*, 1997) mukaan. Klusterin jakaminen voidaan suorittaa esimerkiksi *pääakselia* (principal axis) pitkin (Wu & Zhang, 1991).

Yhdistävien menetelmien tapauksessa on yhdistettävien klustereiden valinta keskeisintä, sillä varsinainen klustereiden yhdistäminen on melkoisen suoraviivaista. Yhdistettävät klusterit voidaan valita esimerkiksi klustereiden lähimpien (single linkage; Sneath & Sokal, 1973) tai kauimpien (complete linkage; King, 1973) alkioiden perusteella. Yhdistettävät klusterit on myös mahdollista valita siten, että klustereiden yhdistäminen heikentää ratkaisua vähiten (Ward's method; Ward, 1963). Hierarkkisista menetelmistä yhdistävät menetelmät ovat yleisimmin käytettyjä ja helpompia toteuttaa kuin jakavat menetelmät (Everitt, 1993; Virmajoki, 2004).

Hierarkkisten menetelmien hyvänä puolena on, että saamme tarvittaessa klusteroinnin kaikille mahdollisille klusterien määrille. Heikkoutena hierarkkisilla algoritmeilla on tietty ahneus, sillä klustereita yhdistellään/jaetaan sen mukaan, mikä juuri kyseisellä hetkellä vaikuttaa parhaalta ratkaisulta. Tehtyjä huonoja päätöksiä ei kuitenkaan pysty myöhemmin perumaan (Kaufman & Rousseeuw, 1990).

4 Osittavat menetelmät

Osittavat menetelmät tuottavat aineistolle vain yhden klusteroinnin (Kaufmann & Rousseeuw, 1990). Osittavat menetelmät jaetaan edelleen optimointimenetelmiin sekä verkkoteoreettisiin, tiheysperustaisiin, ristikkoperustaisiin ja malliperustaisiin menetelmiin.

Optimointimenetelmät jakavat annetun aineiston haluttuun määrään klustereita pyrkien — tapauksesta riippuen — minimoimaan tai maksimoimaan annetun *tavoitefunktion* (objective function) (Jain *& al.*, 1999; Kaufmann & Rousseeuw, 1990; Everitt, 1993). Optimointimenetelmät vaativat siis syötteenään klustereiden lukumäärän, joten klustereiden määrän tulee olla tiedossa tai se on voitava arvioida hyvin etukäteen. Minimoitavana tavoitefunktiona käytetään yleensä keskineliövirhettä. *Keskineliövirhe* (MSE = mean square error) saadaan lasketuksi korottamalla aineiston pisteiden etäisyydet klustereiden keskipisteisiin toiseen potenssiin ja laskemalla niiden keskiarvo. Keskineliövirheen laskeminen on esitetty kaavassa (1), jossa N on aineiston alkioiden lukumäärä, x_i aineiston i:s alkio, c_{p_i} sen klusterin keskipiste, johon alkio x_i kuuluu, ja $d(x_i, c_{p_i})$ on täten aineiston i:nnen alkion ja klusterinsa keskipisteen välinen euklidinen etäisyys.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} d(x_i, c_{p_i})^2 \tag{1}$$

K-means (McQueen, 1967) on eräs tunnetuimpia klusterointimenetelmiä ja se kuuluu optimointimenetelmien

```
1. Valitse satunnaiset alkiot klustereiden keskipisteiksi.
2. TOISTA
3.     Sijoita kukin aineiston alkio siihen klusteriin,
       jonka keskipiste on lähinnä.
4.     Korvaa kunkin klusterin keskipiste klusterin
       alkioiden keskiarvovektorilla.
5. KUNNES (klustereiden keskipisteet eivät enää muutu)
```

Kuva 2: K-means -algoritmi.

luokkaan. K-means käyttää kahta optimaalisuuskriteeriä, joita vuorottelemalla sen onnistuu optimoida lokaalisti kaavan (1) tavoitefunktio. Algoritmi kuvataan useissa lähteissä (Han & Kamber, 2001; Jain *& al.*, 1999; Kaufmann & Rousseeuw, 1990) ja sitä käytetään yleisesti myös toisten menetelmien osana, esimerkiksi Fränti ja Kivijärvi (2000) ovat käyttäneet sitä *paikallishakuun* (local search) pohjautuvan menetelmän osana. K-means -algoritmi on esitetty kuvassa 2.

Verkkoteoreettiset menetelmät muodostavat alkuperäisestä aineistosta verkon ja sen jälkeen ratkaisevat ongelman käyttäen muodostettua verkkoa. Eräs mahdollinen tapa on muodostaa aineistosta *pienin virittävä puu* (minimal spanning tree), jossa solmuina on aineiston alkiot ja kaarten painoina aineiston alkioiden välimatkat (Jain *& al.*, 1999). Tämän jälkeen poistamalla aina painavin kaari voidaan muodostaa klusterointeja eri klusterien lukumäärille, joten verkkoteoreettisilla menetelmillä on selkeitä yhtymäkohtia aiemmin esitettyihin hierarkkisiin menetelmiin.

Tiheysperustaiset menetelmät käsittävät klusteroinnin harvojen alueiden erottamien tiheiden alueiden etsimiseksi (Ma & Chow, 2004). Tiheysperustaisissa menetelmissä aineiston alkiot pyritään jakamaan joukoiksi sellaisia tason alueita, joiden sisältämät alkiot ovat mahdollisimman tiheässä (Han & Kamber, 2001). Han ja Kamber (2001) mainitsevat tiheysperustaisten menetelmien hyvänä puolena, että menetelmät löytävät mielivaltaisen muotoisia ja kokoisia klustereita sekä menetelmien hyvän *hälyn* (noise) siedon. Tiheysperustaisten menetelmien huonona puolena mainitaan, että tiheyden käsitteen määrittely vaatii tuntemusta aineistosta.

Ristikkoperustaiset menetelmät jakavat aineiston alkioiden muodostaman tason äärelliseen määrään neliöitä, jotka yhdessä muodostavat ristikkorakenteen (Han & Kamber, 2001). Tämän jälkeen kaikki klusterointitoimet tehdään ristikkorakenteessa. Han ja Kamber (2001) mainitsevat ristikkoperustaisten menetelmien hyvänä puolena nopean suoritusajan, sillä aikavaativuus on ristikkoperustaisissa menetelmissä tyypillisesti suhteessa solujen määrään, eikä suhteessa aineiston alkioiden määrään. Ma ja Chow (2004) antavat esimerkin tiheys- ja ristikkoperustaisista menetelmistä, sillä heidän esittämänsä menetelmä on yhdistelmä kummastakin menetelmästä.

Malliperustaiset menetelmät pyrkivät hakemaan aineistolle mahdollisimman hyvän selityksen jonkin matemaattisen mallin avulla (Han & Kamber, 2001). Malli voi olla esimerkiksi tilastotieteellinen, jolloin aineisto on *todennäköisyysjakaumien* (propability distributions) tuottamaa, tai *neuraaliverkkoon* (neural network) pohjautuva. Kohosen (1982) esittämä *itseorganisoituva kartta* (self-organizing map) on eräs neuraaliverkkoon pohjautuva malliperustainen menetelmä.

Tässä työssä esitetty klusterointimenetelmien luokitus ei ole mitenkään aukoton, sillä jotkin klusterointimenetelmät ovat yhdistelmiä edellä esitetyistä lähestymistavoista. Tästä esimerkkinä Kaukorannan ja muiden (1998) esittämä *Split-and-merge*-menetelmä, jossa vuoroin jaetaan ja yhdistellään klustereita hierarkkisten menetelmien tapaan, mutta varsinainen menetelmä on pikemminkin optimointimenetelmä. Vastakkaisen esimerkin tarjoaa Likasin ja muiden (2003) esittämä menetelmä, joka on hierarkkinen jakava menetelmä, jossa on käytetty K-meansia menetelmän osana.

Viitteet

Aurenhammer, F. (1991) Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys* **23**(3), 345–405.

Chen, J., Bouman, C.A., Dalton, J.C. (2000) Hierarchical Browsing and Search of Large Image Databases. *IEEE Transactions on Image Processing* **9**(3), 442–455.

Everitt, B.S. (1993) *Cluster Analysis, Third edition*. Arnold, London.

Fränti, P., Gyllenberg, H.G., Gyllenberg, M., Kivijärvi J., Koski T., Lund, T., Nevalainen, O. (2000) Minimizing stochastic complexity using local search and GLA with applications to classification of bacteria. *Biosystems* **57**(1), 37–48.

Fränti, P., Kaukoranta, T., Nevalainen, O. (1997) On the splitting method for VQ codebook generation. *Optical Engineering* **36**(11), 3043–3051.

Fränti, P., Kivijärvi, J. (2000) Randomized Local Search Algorithm for the Clustering Problem. *Pattern Analysis and Applications* **3**(4), 358–369.

Fränti, P., Virmajoki, O. (2004) Optimal clustering by merge-based branch-and-bound. *Submitted for publication 5.10.2004*.

Garey, M.R., Johnson, D.S., Witsenhausen, H.S. (1982) The complexity of the generalized Lloyd-Max problem. *IEEE Transactions on Information Theory* **28**(2), 255–256.

Han, J., Kamber, M. (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco.

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

Kaufman, L., Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Kaukoranta, T., Fränti, P., Nevalainen, O. (1998) Iterative split-and-merge algorithm for VQ codebook generation. *Optical Engineering* **37**(10), 2726–2732.

King, B. (1967) Step-Wise Clustering Procedures. *Journal of the American Statistical Association* **62**(317), 86–101.

Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**(1), 59–69.

Likas, A., Vlassis, N., Verbeek, J.J. (2003) The global k-means clustering algorithm. *Pattern Recognition* **36**(2), 451–461.

Ma, E.W.M., Chow, T.W.S. (2004) A new shifting grid clustering algorithm. *Pattern Recognition* **37**(3), 503–514.

McQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Seal, S., Komarina, S., Aluru, S. (2005) An optimal hierarchical clustering algorithm for gene expression data. *Information Processing Letters* **93**(3), 143–147.

Sneath, P.H.A., Sokal, R.R. (1973) *Numerical Taxonomy*. Freeman, London.

Virmajoki, O. (2004) *Pairwise Nearest Neighbour Method Revisited*. Väitöskirja, Joensuun yliopisto, Joensuu.

Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**(301), 236–244.

Wu, X., Zhang, K. (1991) A Better Tree-Structured Vector Quantizer. *Proceedings Data Compression Conference*, Snowbird, Utah, 392–401.