hyväksymispäivä arvosana

arvostelija

Yksinkertainen alustusalgoritmi k:n keskiarvon ryvästysmenetelmää varten

Panu Luosto

Helsinki 7.2.2008

Seminaarikirjoitelma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Panu Luosto

Työn nimi — Arbetets titel — Title

Yksinkertainen alustusalgoritmi k:n keskiarvon ryvästysmenetelmää varten

Oppiaine — Läroämne — Subject

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Seminaarikirjoitelma	7.2.2008	

Tiivistelmä — Referat — Abstract

Niin sanottu k:n keskiarvon menetelmä (Lloydin menetelmä) on tunnetuimpia ryvästysmenetelmiä. Se on kuitenkin varsin herkkä ensimmäisten ryväskeskusten valinnalle. Yleisesti käytetty tasaisen satunnainen alustus voi johtaa suurella todennäköisyydellä erittäin korkeisiin ryvästyskustannuksiin myös tilanteissa, joissa ryvästysongelma näyttäisi sekä mielekkäältä että helpolta.

Tässä kirjoitelmassa esitellään alustusalgoritmi (D^2-*alustus*), jonka avulla odotusarvoiset ryvästyskustannukset ovat korkeintaan $8(\ln k + 2)$ kertaa niin suuret kuin optimaaliset ryvästyskustannukset. Lausekkeen k tarkoittaa rypäiden lukumäärää. Käytännössä päästään parempiin tuloksiin, koska raja pätee jo alustuksen jälkeisessä tilanteessa, siis ennen itse Lloydin menetelmän suorittamista. Alustusalgoritmi on hyvin yksinkertainen, ja sen aikavaativuus on $O(nkd)$, missä n on ryvästettävien alkioiden lukumäärä ja d dimensio. Kirjoitelman lopussa vertaillaan kokeellisesti tasaisen satunnaisen ja D^2-alustuksen eroja.

Avainsanat — Nyckelord — Keywords

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

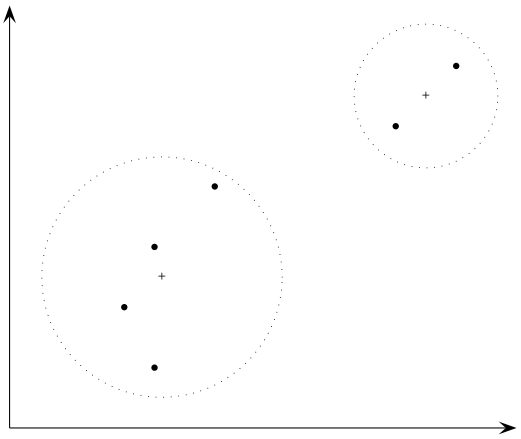
Ryvästys (engl. clustering) on keskeisimpiä tiedon louhinnassa ja ohjaamattomassa oppimisessa käytettyjä menetelmiä. Eri ryvästyslajit poikkeavat toisistaan huomattavasti, mutta tavoitteena on aina annetun joukon sisäisen rakenteen hahmottaminen ilman mitään lisäinformaatiota. Yksinkertaisimmillaan kysymys on joukon osittamisesta siten, että saman osajoukon eli rypään alkiot muistuttavat toisiaan enemmän kuin eri rypäisiin kuuluvat alkiot. Usein ollaan myös kiinnostuneita löytämään ryväskeskukset, eräänlaiset rypäitten edustajat.

Klassisin ryvästysongelma on niin sanottu k:n keskiarvon ryvästys (engl. k-means), jota voidaan pitää euklidisessa avaruudessa tapahtuvan ryvästyksen perustapauksena. Siksi siitä käytetäänkin tässä kirjoitelmassa termiä *euklidinen ryvästys*. Euklidisessa ryvästyksessä on kyse joukon osittamisesta tavalla, joka minimoi tietyn kustannusfunktion arvon. Kunkin rypään keskus on tällöin sen alkioiden keskiarvo. Tähän ryvästysongelmaan liittyy kuuluisa vähittäinen menetelmä, joka esiintyy englanninkielisessä tekstissä niin ikään yleisesti nimellä k-means. Menetelmän esitti tiettävästi ensimmäisenä Lloyd vuonna 1957, vaikka asiaan liittyvä artikkeli julkaistiin vasta 25 vuotta myöhemmin [Llo82]. Termi *Lloydin menetelmä* on verrattaen tunnettu, ja sitä käytetään myös tässä kirjoituksessa, jotta ongelma ja algoritmi voidaan selkeästi erottaa toisistaan. Erityisesti näin halutaan korostaa, että Lloydin menetelmä ei ole euklidisen ryvästysongelman ratkaisualgoritmi eikä tavallisessa muodossaan edes approksimointialgoritmi.

Lloydin menetelmä on erittäin suosittu, koska se on helposti toteutettavissa ja käytännön sovelluksissa myös yleensä nopea. Näennäisestä yksinkertaisuudestaan huolimatta menetelmä on monessa suhteessa varsin huonosti ymmärretty. Sen soveltamiseen liittyy paljon heuristisia käytäntöjä, toisin sanoen tietyissä tilanteissa hyviksi havaittuja ratkaisuja, joiden toimivuudesta ei ole mitään teoreettisia takeita. Tämän kirjoituksen pääsisältönä on esitellä Lloydin menetelmän alustusalgoritmi, jonka avulla ryvästyksen laadulle saadaan tietty odotusarvoinen alajara [AV07]. Algoritmi vaikuttaa lupaavalta myös sovelluksia ajatellen.

2 Euklidinen ryvästys

Tässä luvussa määritellään euklidinen ryvästysongelma. Niin yksi- kuin useampiulotteisten avaruuksien alkioita merkitään jatkossa tavallisilla pienillä kirjaimilla,



Kuva 1: Kuuden tason pisteen optimaalinen euklidinen kaksiryvästys. Ryväskeskukset on merkitty risteillä.

esimerkiksi $x, y \in \mathbb{R}^d$. Tällaisten alkioiden muodostamia joukkoja merkitään isoilla kirjaimilla ($A, B \subset \mathbb{R}^d$) ja joukkojen kokoelmia puolestaan kaunokirjoituskirjaimilla ($\mathcal{R}, \mathcal{S} \subset \mathcal{P}(\mathbb{R}^d)$). Joukon X osituksella tarkoitetaan sellaista X:n epätyhjien osajoukkojen kokoelmaa $\mathcal{R}$, jolla pätee $\bigcup_{A \in \mathcal{R}} A = X$ ja $A \cap B = \emptyset$ kaikilla $A, B \in \mathcal{R}$, $A \neq B$. Tavallisen euklidisen normin merkintänä on käytetty kaksinkertaisia pystyviivoja, joten pisteiden x ja y välinen etäisyys on $d(x, y) = \|x - y\|$.

Euklidisessa ryvästysongelmassa on annettu joukko $X \subset \mathbb{R}^d$ ja haluttu rypäitten lukumäärä $k \in \{1, 2, \ldots, |X|\}$. Ongelmaan liittyvä kustannusfunktio on $\phi_{\mathcal{R}} : \mathcal{R} \to \mathbb{R}$,

$$\phi_{\mathcal{R}}(X) = \sum_{A \in \mathcal{R}} \sum_{x \in A} \|x - c(A)\|^2 \,, \tag{1}$$

missä $c(A) = |A|^{-1} \sum_{x \in A} x$ on joukon A pisteiden keskiarvo. Tehtävänä on määrittää joukon X ositus $\mathcal{R}$ ($|\mathcal{R}| = k$), joka minimoi kustannusfunktion $\phi_{\mathcal{R}}(X)$ arvon. Tällöin ositus $\mathcal{R}$ on joukon X euklidisen k-ryvästysongelman ratkaisu ja ryvästä $A \in \mathcal{R}$ vastava ryväskeskus on $c(A)$. Kuvassa 1 on esitetty tason kuusialkioisen joukon euklidisen kaksiryvästysongelman ratkaisu.

Ongelma esitetään usein muodossa, jossa haetaan k:n keskuksen joukkoa C, joka minimoi kustannukset

$$\phi_C(X) = \sum_{x \in X} \min_{c \in C} \|x - c\|^2 \,. \tag{2}$$

Tällöin keskusta $c \in C$ vastaava ryväs on $\{x \in X \mid k(x) = c\}$, missä $k(x) = \arg\min_{c \in C} \|x - c\|$. Toisin sanoen jokainen piste liitetään siihen rypääseen, jonka

keskus on pistettä lähinnä. Saatava ryvästys ei ole kuitenkaan yksikäsitteinen, koska piste $x \in X$ voi kuulua vain yhteen rypääseen, mutta toisaalta pistettä x lähimpänä olevia keskuksia voi olla useampi kuin yksi. Tästä syystä on selkeämpää samastaa ryvästys ositukseen eikä ryväskeskuksiin, milloin se on mahdollista. Toisaalta monissa tarkasteluissa ollaan kiinnostuneita kustannuksista valittujen keskusten funktiona, ei niinkään itse osituksesta. Tällöin kustannusfunktion muoto (2) on käytännöllisempi.

Luvun 4.2 tarkasteluissa lasketaan useasti vain jonkin mielivaltaisen osajoukon $A \subset X$ osuus ryvästyksen kokonaiskustannuksista. Jos ryvästystä $\mathcal{R}$ vastaa keskusten joukko C, niin tällöin yksinkertaisesti $\phi_{\mathcal{R}}(A) = \phi_C(A) = \sum_{x \in A} \min_{c \in C} \|x - c\|^2$.

Ryvästysongelma on triviaali, jos $k = 1$ tai $k = |X|$. Ensimmäisessä tapauksessa ainoa ryväs sisältää koko joukon X, jälkimmäisessä tapauksessa kuhunkin rypääseen kuuluu täsmälleen yksi piste ja ryvästyskustannukset ovat 0.

Seuraava aputulos on hyödyllinen tarkasteltaessa euklidista ryvästystä.

Aputulos 1 *Olkoon X avaruuden E pistejoukko, jonka alkioiden lukumäärä on n. Olkoon $c(X)$ joukon X pisteiden keskiarvo ja $x \in E$ mielivaltainen piste. Silloin $\sum_{y \in X} \|x - y\|^2 = \sum_{y \in X} \|y - c(X)\|^2 + n\|x - c(X)\|^2$. Lisäksi $\sum_{x,y,\in X} \|x - y\|^2 = 2n \sum_{x \in X} \|x - c(X)\|^2$.*

Todistus. Suoralla laskulla todetaan, että

$$\begin{aligned}
\sum_{y \in X} \|x - y\|^2 &= \sum_{y \in X} \|x - c(X) + c(X) - y\|^2 \\
&= \sum_{y \in X} \|x - c(X)\|^2 + 2 \sum_{y \in X} \langle x - c(X), c(X) - y \rangle + \sum_{y \in X} \|c(X) - y\|^2 \\
&= n\|x - c(X)\|^2 + 2\langle x - c(X), n \cdot c(X) - \sum_{y \in X} y \rangle + \sum_{y \in X} \|y - c(X)\|^2 \, .
\end{aligned}$$

Koska $n \cdot c(X) - \sum_{y \in X} y = n \cdot 1/n \sum_{z \in X} z - \sum_{y \in X} y = 0$, väitteen ensimmäinen osa on tosi. Jälkimmäinen osa saadaan tämän jälkeen helposti:

$$\begin{aligned}
\sum_{x,y \in X} \|x - y\|^2 =& \sum_{x \in X} \left(\sum_{y \in X} \|y - c(X)\|^2 + n\|x - c(X)\|^2 \right) \\
=& n \sum_{y \in X} \|y - c(X)\|^2 + n \sum_{x \in X} \|x - c(X)\|^2 \\
=& 2n \sum_{x \in X} \|x - c(X)\|^2 \, . \quad \square
\end{aligned}$$

Aputuloksen 1 ensimmäisen yhtälön avulla nähdään, että lauseke $\sum_{y \in A} \|x - y\|^2$ on pienimmillään, kun $x = c(A)$. Siis jos yhtälön (1) oikean puolen lausekkeessa korvataan rypään A pisteiden keskiarvo $c(A)$ jollakin muulla pisteellä $c_A \in X$, lausekkeen arvo ei voi pienentyä. Aputuloksen 1 jälkimmäinen yhtälö on hyödyllinen, kun ryvästyksen kustannukset halutaan määrittää laskematta erikseen ryväskeskuksia. Määritelmän yhtälö (1) voidaankin kirjoittaa ilman keskuksia muodossa

$$\phi_{\mathcal{R}}(X) = \sum_{A \in \mathcal{R}} \frac{1}{2|A|} \sum_{x,y \in A} \|x - y\|^2 .$$

Euklidisen ryvästysongelman vaativuuteen liittyy sekaannuksia, joita Aloise et al. ovat käsitelleet [AH07]. Monesti todetaan ongelman olevan NP-kova esittämättä todistusta tai viitteitä, useissa tapauksissa on myös viitattu jonkin muun ryvästysongelman vaativuustodistukseen. Todistus, jonka Drineas et al. ovat esittäneet, on selkeästi virheellinen [DFK+99, AH07]. Tapauksessa, jossa rypäiden lukumäärä k on vähintään 3, voidaan kuitenkin esittää polynomiaalinen palautus kolmiväritettävyysongelmasta [Luo07]. Kolmiväritettävyysongelma on tunnettu NP-täydellinen ongelma [GJ79]. Tapauksen $k = 2$ vaativuus lienee yhä epäselvä.

3 Lloydin menetelmä

Lloydin metodi on kuvailtavissa epämuodollisella tarkkuudella hyvin lyhyesti. Oletetaan, että ryvästettävänä on joukko X ja rypäiden lukumääräksi halutaan k.

1. Valitaan joukosta X keskusten joukko, jonka koko on k.
2. Liitetään kukin piste siihen keskukseen, joka on pistettä lähimpänä. Siirretään sitten kukin keskus siihen liittyvien pisteiden keskiarvokohtaan.
3. Jos näin muodostettu ryvästys on jollakin kriteerillä riittävän hyvä, lopetetaan. Muussa tapauksessa jatketaan jälleen kohdasta 2.

Yleensä ryvästysongelmien yhteydessä puhutaan ryvästettävistä joukoista, niin tehdään tässäkin tutkielmassa. Täsmällisempää olisi kuitenkin puhua monijoukoista, koska käytännön kannalta on epärealistista olettaa, että ryvästettävä aineisto ei koskaan sisältäisi identtisiä alkioita. Koska monijoukkoja on jossakin määrin hankalampaa käsitellä teoreettisesti kuin joukkoja, lienee kuitenkin tarkastelussa monesti helpointa muuntaa monijoukko joukoksi pienten identtisiin alkioihin lisättävien

muutostermien avulla. Tämän jälkeen on tietenkin oltava huolellinen, ettei muutosta käytetä tarkasteluissa liian voimakkaalla tavalla hyväksi. Esimerkiksi kahden alkion monijoukossa $\mathbf{X} = \{\{x, y\}\}$ voi olla $x = y$, mikä ei ole mahdollista muunnetussa joukossa $X = \{x', y'\} = \{x + \delta, y + \epsilon\}$.

Konkreettinen esimerkki monijoukkotarkastelun tarpeellisuudesta on menetelmän ensimmäinen askel eli alustusvaihe. Yleinen käytäntö on arpoa tasaisen satunnaisesti alustavat k keskusta. Tällöin on järkevää varmistua siitä, että kaksi keskusta eivät ole identtiset. Jos ryvästetään joukkoa, riittää identtisyyden varmistamisessa tarkastella alkioiden indeksijoukkoa, mutta monijoukon tapauksessa on vertailtava itse alkioita.

Kohdassa 3 pisteiden ja niiden keskusten välisten neliöllisten etäisyyksien summa ei voi missään vaiheessa kasvaa. Selvästi yksittäisen pisteen osuus kustannukset muodostavasta summasta voidaan minimoida liittämällä se lähimpään keskukseen. Tämä synnyttää joukon uuden osituksen, jossa rypään pisteiden keskiarvo on kustannukset minimoiva ryväskeskus.

Hyvän lopetusehdon määrittäminen ei ole aivan triviaalia. Lloydin menetelmällä löydetään äärellisessä ajassa jokin kustannusfunktion paikallinen minimi, koska erilaisia äärellisen joukon osituksia on äärellinen määrä. Toisaalta pahimmassa tapauksessa tämä voisi viedä liikaa aikaa. Silmukan 3 suorituskerroille onkin syytä asettaa jokin kiinteä yläraja tai vaatia, että ryvästyskustannukset pienentyvät kullakin askeleella riittävän paljon. Käytännön sovelluksissa Lloydin menetelmä stabiloituu yleensä nopeasti, vaikka sen pahimman tapauksen suoritusajan alaraja on ylipolynomiaalinen [AV06].

4 D^2-alustusalgoritmi

Yleisesti käytetty tapa valita alustavat keskukset ryvästettävästä joukosta tasaisen satunnaisesti voi johtaa mielivaltaisen huonoihin ryvästyskustannuksiin. Tämä on ongelma myös realistisissa tilanteissa, kuten luvussa 5 nähdään.

4.1 Algoritmin kuvaus

Seuraavaksi esitettävä alustusalgoritmi, jota kutsutaan tässä kirjoituksessa D^2-alustukseksi, takaa alustuksen jälkeisen ryvästyksen kustannusten odotusarvon ja optimaalisten kustannusten suhteen olevan korkeintaan $8(\ln k + 2)$. Tämä todistetaan

aliluvussa 4.2. Algoritmin ovat esittäneet Arthur et al. [AV07]. He kutsuvat D^2-alustuksen ja Lloydin menetelmän yhdistelmää nimellä *k-means++*.

Olkoot D^2-alustusalgoritmin valitsemat keskukset järjestyksessä $c_1, \ldots, c_k$. Määritellään $D_i(x) = d(x, \{c_1, \ldots, c_i\}) = \min\{\|x - c\| : c \in \{c_1, \ldots, c_i\}\}$. Algoritmi voidaan määritellä seuraavasti.

1. Valitaan keskus c_1 tasaisen satunnaisesti joukosta X.
2. Valitaan järjestyksessä keskukset $c_2, \ldots, c_k$ siten, että alkio $x \in X$ tulee valituksi kullakin kerralla todennäköisyydellä

$$\frac{D_{i-1}(x)^2}{\sum_{y \in X} D_{i-1}(y)^2} \,. \tag{3}$$

Todennäköisyydelle (3) on luonnollinen tulkinta. Tarkoittakoon $\mathcal{R}_i$ ryvästystä, jossa jokainen $x \in X$ on liitetty lähimpään keskukseen joukossa $\{c_1, \ldots, c_i\}$. Silloin kyseinen todennäköisyys voidaan myös kirjoittaa muodossa

$$\frac{\phi_{\mathcal{R}_{i-1}}(\{x\})}{\phi_{\mathcal{R}_{i-1}}(X)} \,. \tag{4}$$

Toisin sanoen kussakin valinnassa tietyn alkion todennäköisyys tulla valituksi on suoraan verrannollinen alkion osuuteen ryvästyskustannuksista, jotka siihen asti valittujen keskusten muodostama ryvästys aiheuttaa. Muodosta (4) nähdään myös, miten alustusalgoritmi on yleistettävissä tapauksiin, joissa käytössä on jokin muu kustannusfunktio.

Jos algoritmin suorituksen aikana pidetään muistissa kunkin alkion kohdalla viitettä sitä lähimpänä olevaan jo valittuun keskukseen, yhden keskuksen valintaan vaadittava aika on luokassa $O(nd)$, missä n on alkioiden lukumäärä ja d avaruuden ulottuvuuksien lukumäärä. Niinpä alustusalgoritmin kokonaisaikavaativuus on $O(ndk)$.

4.2 D^2-alustusalgoritmi on $O(\log k)$-kilpailukykyinen

Arthur et al. osoittavat, että D^2-alustusalgoritmin tuottaman ryvästyksen kustannusten odotusarvo on korkeintaan $8(\ln k + 2)$ kertaa niin suuri kuin optimaalisen ryvästyksen kustannukset [AV07]. Todistus esitetään seuraavassa yksityiskohtaisesti. Huomattakoon, että tulos pätee jo ennen Lloydin algoritmin soveltamista, toisin sanoen tässä alustusalgoritmin valitsemat keskukset määräävät ryvästyksen.

Todistetaan aluksi aputulos, jota tarvitaan myöhemmin kustannusten arviointien yhteydessä.

Aputulos 2 *Kun* $x_i \in \mathbb{R}_+$ *kaikilla* $i \in \{1, \dots m\}$ *ja* $p \in \mathbb{R}$, $p \geq 1$, *niin*

$$\left(\sum_{i=1}^{m} x_i \right)^p \leq m^{p-1} \sum_{i=1}^{m} x_i^p \, .$$

Todistus. Olkoon $\frac{1}{p} + \frac{1}{q} = 1$, missä $p, q \in \mathbb{R}$ ja $p, q \geq 1$. Olkoon lisäksi $x, y \in \mathbb{R}_+^m$, $x = [x_1 \; x_2 \dots x_m]$ ja $y = [m^{-1/q} \; m^{-1/q} \; \dots \; m^{-1/q}]^T$. Hölderin epäyhtälön mukaan $|x \cdot y| \leq \|x\|_p \|y\|_q$. Koska $\|y\|_q = 1$ ja $x_i \geq 0$ kaikilla $i \in \{1, \dots m\}$, niin tämän perusteella

$$\sum_{i=1}^{m} x_i m^{-\frac{1}{q}} \leq \|x\|_p = \left(\sum_{i=1}^{m} x_i^p \right)^{\frac{1}{p}} \, .$$

Epäyhtälön molemmat puolet ovat tässä epänegatiiviset. Tulos saadaan korottamalla epäyhtälö ensin puolittain potenssiin p ja kertomalla se sitten puolittain luvulla $m^{\frac{p}{q}} = m^{p-1}$. □

Seuraava tarkastelu liittyy alustusalgoritmin ensimmäiseen askeleeseen, jossa valitaan yksi ryväskeskus tasaisen satunnaisesti ryvästettävästä joukosta. Vertailukohdaksi kiinnitetään mielivaltainen optimaalinen ryvästys. Valittiin ensimmäinen keskus miten tahansa, se tulee tietenkin valituksi jostakin optimaalisen ryvästyksen rypäästä. Lisäksi kaikkien tämän rypään alkioiden todennäköisyys tulla valituksi on yhtä suuri.

Aputulos 3 *Olkoon* A *optimaalisen ryvästyksen* $\mathcal{R}_{\text{opt}}$ *ryväs ja olkoon* $\mathcal{R}$ *ryvästys, jonka ainoa keskus on valittu tasaisen satunnaisesti joukosta* A. *Tällöin* $\mathbb{E}[\phi_{\mathcal{R}}(A)] = 2\phi_{\text{opt}}(A)$.

Todistus. Käyttämällä hyväksi aputuloksen 1 jälkimmäistä kohtaa nähdään, että

$$\begin{aligned} \mathbb{E}[\phi_{\mathcal{R}}(A)] &= \sum_{y \in A} \frac{1}{|A|} \sum_{x \in A} \|x - y\|^2 \\ &= \frac{1}{|A|} 2|A| \sum_{x \in A} \|x - c(A)\|^2 \\ &= 2 \sum_{x \in A} \|x - c(A)\|^2 \, . \end{aligned}$$

Koska $\mathcal{R}_{\text{opt}}$ on optimaalinen ryvästys, siinä joukon A ryväskeskus on A:n alkioiden keskiarvo. Väite tuli siten todistetuksi. □

Arvioidaan seuraavaksi ryvästyskustannuksia tilanteessa, jossa yksittäinen uusi keskus valitaan D^2-painotusta käyttäen. Tällöin on olemassa jokin ryvästys $\mathcal{R}$, johon uusi keskus lisätään. Seuraavassa ei oleteta mitään siitä, mikä on ryvästyksen $\mathcal{R}$ keskusten joukon ja optimaalisen ryvästyksen alkion A leikkauksen koko.

Aputulos 4 *Olkoon A optimaalisen ryvästyksen $\mathcal{R}_{\text{opt}}$ ryväs ja $\mathcal{R}$ ryvästys, joka sisältää vähintään yhden keskuksen. Laaditaan ryvästys $\mathcal{R}'$ lisäämällä ryvästykseen $\mathcal{R}$ uusi keskus joukosta A käyttäen D^2-alustusalgoritmia. Tällöin pätee $\mathbb{E}[\phi_{\mathcal{R}'}(A)] \leq 8\phi_{\text{opt}}(A)$.*

Todistus. Merkitään seuraavassa pistettä x lähinnä olevaa ryvästyksen $\mathcal{R}$ keskusta $k(x)$. Siis $k(x) = \arg\min_{y \in C} \|x-y\|$, missä C on ryvästyksen $\mathcal{R}$ keskusten joukko. Jos lähimpiä keskuksia on useampi kuin yksi, voidaan keskus $k(x)$ kiinnittää mielivaltaisesti. Olkoon lisäksi pisteen x etäisyys lähimpään $\mathcal{R}$:n keskukseen $D(x) = \|x-k(x)\|$. Kun tiedetään pisteen c tulevan valituksi joukosta A, todennäköisyys valita c on

$$\mathbb{P}_c = \frac{D(c)^2}{\sum_{x \in A} D(x)^2} \, .$$

Kun keskus c on valittu, pisteen $y \in A$ osuus ryvästyksen $\mathcal{R}'$ kustannuksista on $\min\{D(y)^2, \|y-c\|^2\}$. Siis

$$\begin{aligned} \mathbb{E}[\phi_{\mathcal{R}'}(A)] &= \sum_{c \in A} \mathbb{P}_c \sum_{y \in A} \min\{D(y)^2, \|y-c\|^2\} \\ &= \sum_{c \in A} \frac{D(c)^2}{\sum_{x \in A} D(x)^2} \sum_{y \in A} \min\{D(y)^2, \|y-c\|^2\} \, . \end{aligned} \tag{5}$$

Kolmioepäyhtälöä hyväksi käyttäen saadaan arvio $D(c) = \|c-k(c)\| \leq \|c-k(z)\| \leq \|c-z\| + \|z-k(z)\| = \|z-c\| + D(z)$, missä $z \in A$. Aputuloksen 2 perusteella siten $D(c)^2 \leq 2D(z)^2 + 2\|z-c\|^2$. Summaamalla kaikkien $z \in A$ yli ja jakamalla joukon A alkioiden lukumäärällä saadaan

$$D(c)^2 \leq \frac{2}{|A|} \sum_{z \in A} D(z)^2 + \frac{2}{|A|} \sum_{z \in A} \|z-c\|^2 \, .$$

Sijoitetaan tämä epäyhtälöön 5, jolloin voidaan lopulta arvioida:

$$\begin{aligned}
\mathbb{E}[\phi_{\mathcal{R}'}(A)] &\leq \frac{2}{A}\sum_{c\in A}\frac{\sum_{z\in A} D(z)^2}{\sum_{x\in A} D(x)^2}\sum_{y\in A}\min\{D(y)^2, \|y-c\|^2\} \\
&\quad + \frac{2}{|A|}\sum_{c\in A}\frac{\sum_{z\in A}\|z-c\|^2}{\sum_{x\in A} D(x)^2}\sum_{y\in A}\min\{D(y)^2, \|y-c\|^2\} \\
&\leq \frac{2}{A}\sum_{c\in A}\frac{\sum_{z\in A} D(z)^2}{\sum_{x\in A} D(x)^2}\sum_{y\in A}\|y-c\|^2 + \frac{2}{|A|}\sum_{c\in A}\frac{\sum_{z\in A}\|z-c\|^2}{\sum_{x\in A} D(x)^2}\sum_{y\in A} D(y)^2 \\
&= \frac{4}{|A|}\sum_{c\in A}\sum_{y\in A}\|y-c\|^2 \\
&= 4\cdot 2\phi_{\text{opt}}(A) \qquad (6) \\
&= 8\phi_{\text{opt}}(A)\ .
\end{aligned}$$

Kohta (6) saatiin aputuloksesta 3. □

Äskeisten aputulosten perusteella tiedetään, että D^2-alustuksen tuottamien ryvästyskustannuksien odotusarvo olisi vakiokerrointa vailla optimaalinen, mikäli mitkään kaksi valittua keskusta eivät kuuluisi samaan optimaalisen ryvästyksen rypääseen. Tähän ei kuitenkaan päästä. Väite alustusalgoritmin $O(\log k)$-kilpailukykyisyydestä pystytään todistamaan soveltamalla seuraavaa aputulosta tilanteessa, jossa on valittu vain yksi keskus optimaalisen ryvästyksen rypäästä A.

Aputulos 5 *Olkoon $\mathcal{R}$ joukon X ryvästys ja olkoon $\mathcal{R}_{\text{opt}}$ X:n optimaalinen ryvästys. Valitaan osajoukko $\mathcal{U} \subset \mathcal{R}_{\text{opt}}$ siten, että $|\mathcal{U}| = u > 0$. Olkoon $U = \bigcup_{B\in\mathcal{U}} B$ näiden rypäitten sisältämien pisteiden joukko ja $V = X \setminus U$. Muodostetaan ryvästys $\mathcal{R}'$ lisäämällä $t \leq u$ satunnaista keskusta ryvästykseen $\mathcal{R}$ käyttäen D^2-alustusalgoritmin valintamenettelyä. Saadun ryvästyksen kustannusten odotusarvolle saadaan yläraja*

$$\mathbb{E}[\phi_{\mathcal{R}'}(X)] \leq (1+H_t)\left(\phi_{\mathcal{R}}(V) + 8\phi_{\text{opt}}(U)\right) + \frac{u-t}{u}\,\phi_{\mathcal{R}}(U)\ ,$$

missä $H_t = \sum_{i=1}^{t} 1/t$. Sovitaan tässä, että $H_0 = 0$.

Todistus. Todistetaan väite induktiolla jonon

$$(a_n) = ((t_n, u_n)) = ((0,1),(1,1),(0,2),(1,2),(2,2),\dots)$$

yli. Käsitellään ensin perustapaukset $(t,u) \in \{(0,u),(1,1) \mid u \in \mathbb{N}\}$. Merkitään seuraavassa luettavuuden parantamiseksi $\phi \equiv \phi_{\mathcal{R}}$.

Kun $t = 0$ ja $u > 0$, ryvästys ei muutu, eli $\mathcal{R}' = \mathcal{R}$. Saadaan triviaali arvio

$$\begin{aligned}\mathbb{E}[\phi_{\mathcal{R}'}(X)] &= \phi(X) \\ &\leq \phi(X) + 8\phi_{\text{opt}}(U) \\ &= (1+0) \cdot (\phi(V) + 8\phi_{\text{opt}}(U)) + 1 \cdot \phi(U) \ .\end{aligned}$$

Olkoon nyt $t = u = 1$. On siis valittu optimaalisesta ryvästyksestä yksi ryväs. Todennäköisyys, että ryvästykseen $\mathcal{R}$ lisättävä keskus on joukosta U, on $\phi(U)/\phi(X)$. Aputuloksen 4 mukaan siinä tapauksessa $\mathbb{E}[\phi_{\mathcal{R}'}(U)] \leq 8\phi_{\text{opt}}(U)$, ja koska joukon V ryvästämisen kustannukset eivät ainakaan kasva, pätee $\mathbb{E}[\phi_{\mathcal{R}'}(X)] \leq \phi(V) + 8\phi_{\text{opt}}(U)$. Luonnollisesti siinäkään tapauksessa, että uusi keskus valitaan joukosta V, ryvästyksen kulut eivät voi kasvaa. Näin pystytään arvioimaan

$$\begin{aligned}\mathbb{E}[\phi_{\mathcal{R}'}(X)] &\leq \frac{\phi(U)}{\phi(X)} \left(\phi(V) + 8\phi_{\text{opt}}(U)\right) + \frac{\phi(V)}{\phi(X)}\phi(X) \\ &\leq \phi(V) + 8\phi_{\text{opt}}(U) + \phi(V) \qquad (7) \\ &= 2\phi(V) + 8\phi_{\text{opt}}(U) \\ &\leq (1+1) \cdot (\phi(V) + 8\phi_{\text{opt}}(U)) + 0 \cdot \phi(U) \ .\end{aligned}$$

Kohta (7) seurasi siitä, että $\phi(U)/\phi(X) \leq 1$.

Oletetaan nyt väitteen pitävän paikkansa arvoilla $(t-1, u)$ ja $(t-1, u-1)$. Näytetään, että tällöin väite on tosi myös arvoilla (t, u). Tämän tuloksen avulla voidaan induktiotodistus viedä helposti päätökseen.

Jaetaan käsittely kahteen osaan sen mukaan, kuuluuko ensimmäinen ryvästykseen $\mathcal{R}$ lisättävä uusi ryväskeskus joukkoon V vai U. Vastaavien tapahtumien todennäköisyydet ovat $\phi(V)/\phi(X)$ ja $\phi(U)/\phi(X)$.

Ensimmäisessä tapauksessa syntyy ryvästys $\mathcal{T}$, joka sisältää yhden keskuksen enemmän kuin $\mathcal{R}$. Tietenkin $\phi_{\mathcal{T}}(A) \leq \phi_{\mathcal{R}}(A)$ kaikilla $A \subset X$. Käytetään tätä tietoa hyödyksi ja sovelletaan induktio-oletusta ryvästykseen $\mathcal{T}$ arvoilla $t-1$ ja u. Ryvästyksen $\mathcal{R}'$ odotusarvoisten kustannuksien ylärajaksi saadaan tässä tapauksessa

$$\mathbb{E}[\phi_{\mathcal{R}'}(X)|\mathcal{T}] \leq (1 + H_{t-1}) \cdot (\phi(V) + 8\phi_{\text{opt}}(U)) + \frac{u-(t-1)}{u}\ \phi(U) \ .$$

Vastaavasti tapahtuman osuus koko kustannusten $\phi_{\mathcal{R}'}(X)$ odotusarvosta on

$$\frac{\phi(V)}{\phi(X)} \left((1 + H_{t-1}) \cdot (\phi(V) + 8\phi_{\text{opt}}(U)) + \frac{u-t+1}{u}\phi(U) \right) \ . \qquad (8)$$

Tarkastellaan nyt toista tapausta, jossa ensimmäinen uusi keskus a tulee valituksi jostakin optimaalisen ryvästyksen rypäästä $A \in \mathcal{U}$. Kiinnitetään jokin $a \in A$, ja olkoon näin syntynyt ryvästys $\mathcal{U}_a$. Olkoon p_a todennäköisyys valita alkio a keskukseksi sillä ehdolla, että keskus valitaan joukosta A. Merkitään lisäksi $U' = U \setminus A$ ja $V' = V \cup A$. Nytkin luonnollisesti $\phi_{\mathcal{U}_a}(A) \leq \phi(A)$ kaikilla $A \subset X$. Sovelletaan induktio-oletusta ryvästykseen $\mathcal{U}_a$ arvoilla $(t-1, u-1)$, jolloin voidaan arvioida ryvästyksen $\mathcal{R}'$ kustannusten odotusarvon olevan korkeintaan

$$
\begin{aligned}
&\mathbb{E}[\phi_{\mathcal{R}'}(X)|\mathcal{U}_a] \\
&\quad\leq \sum_{a\in A} p_a \left((1+H_{t-1}) \left(\phi_{\mathcal{U}_a}(V') + 8\phi_{\text{opt}}(U')\right) + \frac{(u-1)-(t-1)}{u-1}\phi_{\mathcal{U}_a}(U') \right) \\
&\quad\leq \sum_{a\in A} p_a \left((1+H_{t-1}) \Big(\phi_{\mathcal{U}_a}(V') + 8\phi_{\text{opt}}(U) - 8\phi_{\text{opt}}(A)\Big) + \frac{u-t}{u-1}\phi(U') \right) \\
&\quad\leq \sum_{a\in A} p_a \Bigg((1+H_{t-1}) \Big(\phi_{\mathcal{U}_a}(V) + \phi_{\mathcal{U}_a}(A) + 8\phi_{\text{opt}}(U) - 8\phi_{\text{opt}}(A)\Big) \\
&\qquad\quad + \frac{u-t}{u-1}(\phi(U) - \phi(A)) \Bigg) .
\end{aligned}
$$

Aputuloksen 4 mukaan $\sum_{a\in A} p_a \phi_{\mathcal{U}_a}(A) \leq 8\phi_{\text{opt}}(A)$. Koska lisäksi $\sum_{a\in A} p_a = 1$, ja $\phi_{\mathcal{U}_a}(V) \leq \phi(V)$, niin yllä olevasta seuraa edelleen

$$
\mathbb{E}[\phi_{\mathcal{R}'}(X)|\mathcal{U}_a] \leq (1+H_{t-1})(\phi(V) + 8\phi_{\text{opt}}(U)) + \frac{u-t}{u-1}(\phi(U) - \phi(A)) .
$$

Nyt voidaan viedä toisen osatapauksen tarkastelu päätökseen. On tarkasteltu siis tilannetta, jossa ensimmäinen valittava keskus kuuluu johonkin optimaalisen ryvästyksen rypääseen $A \in \mathcal{U}$. Tällöin syntyvien kustannusten osuus koko odotusarvosta $\mathbb{E}[\phi_{\mathcal{R}'}(X)]$ on korkeintaan

$$
\begin{aligned}
&\sum_{A\in\mathcal{U}} \frac{\phi(A)}{\phi(X)} \left((1+H_{t-1})(\phi(V) + 8\phi_{\text{opt}}(U)) + \frac{u-t}{u-1}(\phi(U) - \phi(A)) \right) \\
&\quad= (1+H_{t-1}) \left(\phi(V) + 8\phi_{\text{opt}}(U)\right) \frac{1}{\phi(X)} \sum_{A\in\mathcal{U}} \phi(A) \\
&\qquad + \frac{u-t}{u-1} \cdot \frac{1}{\phi(X)} \sum_{A\in\mathcal{U}} \left(\phi(A)\phi(U) - \phi(A)^2\right) \qquad (9) \\
&\quad\leq (1+H_{t-1}) \left(\phi(V) + 8\phi_{\text{opt}}(U)\right) \frac{\phi(U)}{\phi(X)} \;+\; \frac{u-t}{u-1} \cdot \frac{1}{\phi(X)} \left(\phi(U)^2 - \frac{1}{u}\phi(U)^2 \right) \\
&\quad= \frac{\phi(U)}{\phi(X)} \left((1+H_{t-1})(\phi(V) + 8\phi_{\text{opt}}(U)) + \frac{u-t}{u}\phi(U) \right) . \qquad (10)
\end{aligned}
$$

Kohdassa (9) käytettiin arviota, joka seuraa suoraan aputuloksesta 2:

$$\sum_{A\in\mathcal{U}} \phi(A)^2 \geq \frac{1}{u}\left(\sum_{A\in\mathcal{U}} \phi(A)\right)^2 = \frac{1}{u}\phi(U)^2 \,.$$

Yhdistämällä kaksi edellä käsiteltyä osatapausta ja niitä vastaavat arviot (8) ja (10) voidaan näyttää, että aputuloksen 5 ollessa voimassa arvoilla $(t-1, u)$ ja $(t-1, u-1)$, se on voimassa myös arvoilla (t, u):

$$\begin{aligned}
\mathbb{E}[\phi_{\mathcal{R}'}(X)] &\leq \frac{\phi(V)}{\phi(X)}\left((1+H_{t-1})(\phi(V)+8\phi_{\text{opt}}(U)) + \frac{u-t-1}{u}\phi(U)\right) \\
&\quad + \frac{\phi(U)}{\phi(X)}\left((1+H_{t-1})(\phi(V)+8\phi_{\text{opt}}(U)) + \frac{u-t}{u}\phi(U)\right) \\
&= \frac{\phi(V)+\phi(U)}{\phi(X)}(1+H_{t-1})\left(\phi(V)+8\phi_{\text{opt}}(U)\right) \\
&\quad + \frac{\phi(U)}{u\cdot\phi(X)}\Big(\phi(V)(u-t+1)+\phi(U)(u-t)\Big) \\
&= (1-H_{t-1})\left(\phi(V)+8\phi_{\text{opt}}(U)\right) \qquad (11) \\
&\quad + \frac{\phi(U)}{u\cdot\phi(X)}\Big(\phi(V)(u-t+1)+(\phi(X)-\phi(V))(u-t)\Big) \\
&= (1-H_{t-1})\left(\phi(V)+8\phi_{\text{opt}}(U)\right) \\
&\quad + \frac{\phi(U)}{u\cdot\phi(X)}\big(\phi(V)+\phi(X)u-\phi(X)t\big) \\
&= (1+H_{t-1})\left(\phi(V)+8\phi_{\text{opt}}(U)\right) + \frac{\phi(U)}{\phi(X)}\frac{\phi(V)}{u} + \frac{u-t}{u}\phi(U) \\
&\leq \left(1+H_{t-1}+\frac{1}{u}\right)\big(\phi(V)+8\phi_{\text{opt}}(U)\big) + \frac{u-t}{u}\phi(U) \qquad (12) \\
&\leq (1+H_t)\left(\phi(V)+8\phi_{\text{opt}}(U)\right) + \frac{u-t}{u}\phi(U) \,.
\end{aligned}$$

Kohta (11) seuraa siitä, että $\phi(V)+\phi(U) = \phi(X)$, ja kohta (12) siitä, että $\phi(U)/\phi(X) \leq 1$. Viimeinen epäyhtälö saadaan, koska $t \leq u$.

Nyt todistus voidaan viedä päätökseen. Oletetaan aputuloksen 5 väitteen olevan tosi jonon $(a_n) = (a_1, a_2, a_3 \dots)$ indekseillä $k \geq 3$. Olkoon $a_{k+1} = (t, u)$. Väitteen

osoitettiin pätevän, jos $t = 0$. Muussa tapauksessa induktio-oletuksen perusteella väite pätee arvopareilla $(t-1, u)$ ja $(t-1, u-1)$, joten äskeisen tarkastelun perusteella se pätee myös tapauksessa (t, u). □

Lause 6 *Jos joukosta X valitaan k keskusta käyttäen D^2-alustusalgoritmia, saadun ryvästyksen kustannusten odotusarvolle pätee* $\mathbb{E}[\phi(X)] \leq 8(\ln k + 2)\phi_{\text{opt}}(X)$.

Todistus. Kiinnitetään jokin joukon X optimaalinen k-ryvästys. Valitkoon algoritmi ensimmäisen keskuksen tämän optimaalisen ryvästyksen rypäästä A. Muodostetaan ryvästys $\mathcal{R}'$, joka sisältää tämän yhden keskuksen. Aputuloksen 3 mukaan $\mathbb{E}[\phi_{\mathcal{R}'}(A)] = 2\phi_{\text{opt}}(A)$. Kun algoritmi valitsee tämän jälkeen puuttuvat $k-1$ keskusta, voidaan aputulosta 5 soveltaa arvoilla $U = X \setminus A$ ja $t = u = k - 1$. Siten

$$\begin{aligned}\mathbb{E}[\phi(X)] &\leq (1 + H_{k-1})\big(\phi_{\mathcal{R}'}(A) + 8\phi_{\text{opt}}(X \setminus A)\big) + \frac{0}{u}\phi_{\mathcal{R}'}(X \setminus A) \\ &\leq (1 + H_{k-1})\big(\phi_{\mathcal{R}'}(A) + 8\phi_{\text{opt}}(X) - 8\phi_{\text{opt}}(A)\big)\ ,\end{aligned}$$

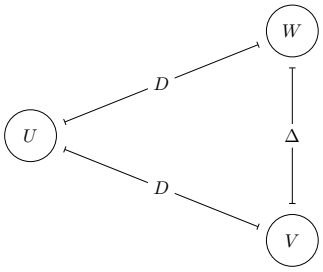
mistä seuraa

$$\begin{aligned}\mathbb{E}[\phi(X)] &\leq (1 + H_{k-1})\big(2\phi_{\text{opt}}(A) + 8\phi_{\text{opt}}(X) - 8\phi_{\text{opt}}(A)\big) \\ &\leq (1 + \ln k + 1)\, 8\phi_{\text{opt}}(X) \\ &= 8(\ln k + 2)\phi_{\text{opt}}(X)\ . \quad \square\end{aligned}$$

Äskeinen tulos D^2-alustuksen tuottaman ryvästyksen odotusarvon $O(\ln k)$-kilpailukykyisyydestä optimaaliseen ryvästyksen nähden on vakiotekijää vaille tiukka. Voidaan nimittäin konstruoida säännöllinen joukko, jonka optimaalisessa ryvästyksessä rypäät ovat hyvin tiiviitä verrattuna rypäiden välisiin etäisyyksiin. Raja-arvona saadaan tulos, jonka mukaan D^2-ryvästyksen kulujen odotusarvo on ainakin $2 \ln k$ kertaa niin suuri kuin optimaalisessa tapauksessa [AV07].

5 Kilpailukykyisyystuloksen tarkastelua

Kilpailukykyisyystulos herättää välittömästi lisätarkastelua ansaitsevia kysymyksiä. Kuinka paljon itse asiassa tulos kertoo meille ryvästyksen laadusta? Mikä on onnistuneen alustuksen merkitys sen jälkeen suoritettavalle Lloydin menetelmälle?



Kuva 2: Kolme tiheää ryvästä.

On helppoa konstruoida yksinkertaisia kaksiryvästysongelmia, joissa johdettu kustannusten yläraja on hyvin löysä. Toisaalta kaukana toisistaan sijaitsevien tiheiden rypäiden sommitelmat ovat mielenkiintoisia alustusmenetelmien vertailua ajatellen.

Kuva 2 havainnollistaa tilannetta, jossa tasaisen satunnaisen alustuksen käyttäminen johtaa mielivaltaisen korkeaan ryvästyskustannusten odotusarvoon. Oletetaan, että kahden eri alkion välinen etäisyys kunkin joukon U, V tai W sisällä on δ ja että eri joukkoihin kuuluvien alkioiden välinen etäisyys on kuvan mukaisesti aina joko täsmälleen Δ tai D. Lisäksi olkoon $\delta < \Delta < D$. Jos alustavat kolme keskusta poimitaan tasaisen satunnaisesti, todennäköisyys valita täsmälleen kaksi keskusta joukosta U on $2/9$. Koska oletuksena on $\Delta < D$, kumpikaan joukon U alustavista keskuksista ei pääse koskaan siirtymään joukkoon $V \cup W$. Antamalla välimatkojen Δ ja D kasvaa rajatta, oletusarvoiset ryvästyskustannukset kasvavat myös rajatta.

Vastaavanlaisia tilanteita syntyy helposti, mikäli optimaaliset rypäät ovat riittävän tiheitä verrattuna rypäitten välisiin etäisyyksiin. Vaatimuksena on oikeastaan vain, että tilanne ei ole täysin symmetrinen. Tämän kirjoittaja generoi testimateriaaliksi 10 000 pistettä 100-ulotteisesta hyperkuutiosta, jonka sivun pituus oli 100. Joukosta $\{0, 1, \ldots, 100\}^{100}$ arvottiin tasaisen satunnaisesti 100 alkiota, joita käytettiin 100 arvottavan 100-alkioisen gaussisen joukon odotusarvoina. Kunkin joukon kovarianssimatriisi oli ykkösmatriisi. Tällaisessa tilanteessa arvottavat pisteet muodostavat hyvin todennäköisesti 100 kaukana toisistaan sijaitsevaa tiheää ryvästä. Aineisto ryvästettiin Lloydin menetelmällä käyttäen 100 kertaa alustusalgoritmina D^2-alustusta ja 100 kertaa tasaisen satunnaista alustusta. Optimaalista ryvästystä approksimoitiin käyttämällä ryväskeskuksina niitä Gaussin jakaumien odotusarvoja, joiden avulla pisteet oli generoitu. Myös pelkän D^2-alustuksen tuottaman ryvästyksen kulut määritettiin.

Tässä tilanteessa Lloydin menetelmän onnistuminen on täysin riippuvainen alustuk-

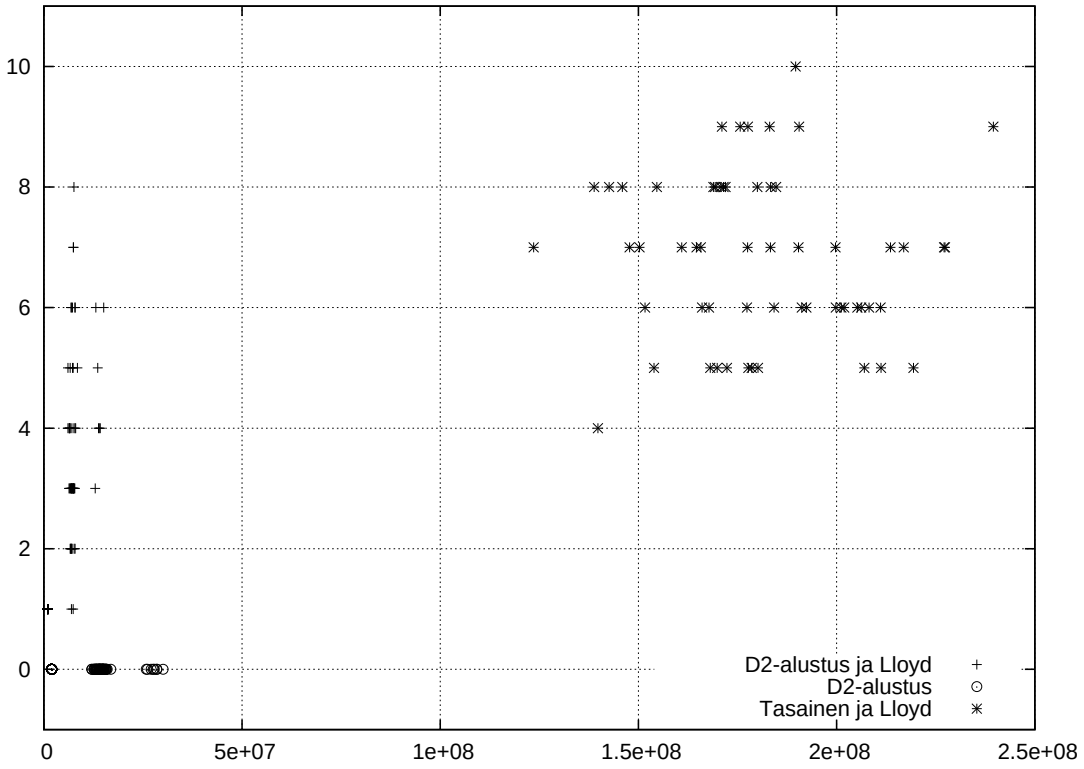
sen onnistumisesta. Kuvassa 3 ovat vaaka-akselilla lopulliset ryvästyskustannukset ja pystyakselilla Lloydin menetelmän vakautumiseen tarvittavien silmukka-askelien lukumäärä eri ryvästyskerroilla. D^2-alustusta käyttävä Lloydin menetelmä löysi optimaalisen ryvästyksen approksimaatiotuloksen 42 % tapauksista, näitä vastaa kuvan vasemmanpuoleisin risti. Varsin hyviä ryvästyksiä tuotti myös pelkän D^2-alustuksen käyttäminen ilman Lloydin menetelmää (ympyrät suoralla $y = 0$). Sen sijaan tasaisen satunnaisen alustuksen ja Lloydin menetelmän yhdistelmä oli koetilanteessa aina molempia edellä mainittuja menetelmiä merkittävästi huonompi. Taulukossa 1 on muutamia kokeeseen liittyviä tunnuslukuja.

Äskeinen ryvästysongelma oli hyvin selkeä, koska välimatkat optimaalisen ryvästyksen rypäiden välillä olivat suuret verrattuna etäisyyksiin kunkin rypään sisällä. Tasaisen satunnainen alustus valitsi kuitenkin suurella todennäköisyydellä useita alustavia keskuksia saman tiheän rypään sisältä, ja tilanteen epäsymmetrisyyden vuoksi Lloydin menetelmä ei pystynyt enää siirtämään kaikkia keskuksia omiin rypäisiinsä. Arhur et al. ovat myös suorittaneet suppean joukon empiirisiä testejä, joiden tulokset ovat samansuuntaisia kuin tässä esitetty [AV07].

6 Lopuksi

Lloydin menetelmän ongelmiin kuuluu sen herkkyys käytetylle alustukselle eli ensimmäisten keskusten valinnalle. Kirjoituksessa esiteltiin yksinkertainen alustusalgoritmi, joka takaa ryvästyskustannusten odotusarvolle $O(\ln k)$-ylärajan optimaalisiin kustannuksiin nähden. Empiirisessä kokeessa nähtiin tasaisen satunnaisen alustuksen johtavan suuriin ryvästyskustannuksiin tilanteessa, jossa tiheät ja säännölliset gaussiset rypäät olivat selkeästi erillään toisistaan. D^2-alustus tuotti tässä tilanteessa selkeästi parempia tuloksia.

Kiinnostuneen lukijan on helppoa löytää runsaasti Lloydin menetelmään liittyviä artikkeleita. Oppikirjoista mainittakoon Bishopin ja Alpaydinin teokset [Bis06] [Alp04], joissa molemmissa käsitellään muun muassa menetelmän yhteyksiä Gaussin mikstuurimalleihin ja odotusarvon maksimointialgoritmiin (EM-algoritmi). Euklidiseen ryvästysongelmaan on kehitetty lukuisia approksimaatioalgoritmeja, joiden käytännöllinen merkitys on hyvin vähäinen. Ostrovsky et al. esittelevät kuitenkin algoritmin, joka nivoutuu kiinnostavalla tavalla D^2-alustusalgoritmiin [ORSS06].



Kuva 3: Ryvästyskustannukset testitapauksessa. Vaaka-akselilla ovat ryvästyskustannukset ja pystyakselilla Lloydin menetelmän vakautumiseen vaadittavien toistojen määrä. Kun ryvästys muodostetaan suoraan D^2-alustuksen tuottamien keskusten perusteella, Lloydin menetelmää ei ajeta ollenkaan (ympyrät suoralla $y = 0$).

	D^2 **ja Lloyd**	D^2	**Tasainen ja Lloyd**
minimi	1,00	1,97	125
keskiarvo	5,07	10,3	183
maksimi	15,2	30,4	242
keskihajonta	3,89	7,96	24,7

Taulukko 1: Ryvästyskustannukset testitapauksessa. Optimaalisen ryvästyksen approksimaation tuottamiksi kuluiksi on tässä asetettu 1. D^2 *ja Lloyd* tarkoittaa D^2-alustuksen ja Lloydin menetelmän yhdistelmää, ja *Tasainen ja Lloyd* tasaisen satunnaisen alustuksen ja Lloydin menetelmän yhdistelmää. Pelkkä D^2 viittaa ryvästyskustannuksiin, kun ryväskeskuksina käytetään suoraan D^2-alustuksen tuottamia keskuksia.

Lähteet

AH07 Aloise, D. ja Hansen, P., On the complexity of minimum sum-of-squares clustering. Tekninen raportti G-2007-50, Les Cahiers du GERAD, Montréal (Québec) Canada, July 2007.

Alp04 Alpaydin, E., *Introduction To Machine Learning.* MIT Press, 2004.

AV06 Arthur, D. ja Vassilvitskii, S., How slow is the k-means method? *SCG '06: Proceedings of the twenty-second annual symposium on Computational geometry*, New York, NY, USA, 2006, ACM Press, sivut 144–153.

AV07 Arthur, D. ja Vassilvitskii, S., k-means++: the advantages of careful seeding. *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 2007, Society for Industrial and Applied Mathematics, sivut 1027–1035.

Bis06 Bishop, C. M., *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer, 2006.

DFK+99 Drineas, P., Frieze, A., Kannan, R., Vempala, S. ja Vinay, V., Clustering in large graphs and matrices. *SODA '99: Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 1999, Society for Industrial and Applied Mathematics, sivut 291–299.

GJ79 Garey, M. R. ja Johnson, D. S., *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., New York, NY, USA, 1979.

Llo82 Lloyd, S., Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28,2(Mar 1982), sivut 129–137.

Luo07 Luosto, K., keskustelu, December 2007.

ORSS06 Ostrovsky, R., Rabani, Y., Schulman, L. J. ja Swamy, C., The effectiveness of lloyd-type methods for the k-means problem. *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 47, sivut 165–176.