Tiedon louhintaa kompressointia hyväksikäyttäen

Toni Merivuori

Helsinki 3. maaliskuuta 2008

Tiedon louhinnan seminaari, kevät 2008

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Kaikelle bittijonoina esitettävälle tiedolle on ominaista merkkijonojen algoritmiset ominaisuudet, kuten pakkautuvuus. Objektin Kolmogorov-kompleksisuus kuvaa lyhimmän ohjelman pituuden, joka sellaisenaan muodostaa objektin. Kolmogorov-kompleksisuuden laskeminen on ratkeamaton ongelma, mutta se antaa teoreettisen alarajan mille tahansa pakkausohjelmalle. Normalisoitu informaatioetäisyys perustuu Kolmogorov-kompleksisuuteen. Se esittää teoreettisen kehyksen kahden objektin samankaltaisuuden tarkasteluun. Normalisoitu kompressioetäisyys on edellisen approksimaatio. Siinä teoreettinen Kolmogorov-kompleksisuus korvataan normaalikompressorilla, joka voi olla lähes mikä tahansa pakkausohjelma. Nyt voidaan käytännössä mitata kahden objektin algoritmista samankaltaisuutta. Objektijoukon kesken voidaan mitata kaikkien joukon alkioiden väliset keskinäisetäisyydet ja muodostaa etäisyysmatriisi. Kvartetti-menetelmä ryvästää hyvin samankaltaiset objektit omiin rypäisiinsä. Menetelmän universaalisen luonteen ansiosta sitä voidaan sellaisenaan soveltaa hyvin moniin tiedon louhinnan ongelmiin.

Sisältö

1 Johdanto

Seminaarityö esittelee datan kompressointiin perustuvan, universaaliksikin kutsutun menetelmän kahden objektin samankaltaisuuden mittaamiseen sekä Kvartetti-menetelmän luokitellun tiedon ryvästämiseen. Menetelmää voidaan soveltaa moniin tiedon louhinnan ongelmiin sellaisenaan. Sen avulla vertaillaan kahta objektia perustuen kaikelle datalle yhteiseen universaaliseen ominaisuuteen, pakkautuvuuteen. Vertailun tulokset voidaan esittää etäisyysmatriisissa, josta dataa voidaan edelleenkäsitellä esimerkiksi ryvästämällä.

Toinen luku esittelee perusteet algoritmisesta informaatioteoriasta, jonka avulla määritellään kahden objektin välinen informaatioetäisyys. Normalisoimalla informaatioetäisyys saadaan hyvin määritelty etäisyysmitta kahden objektin samankaltaisuuden tutkimiseen. Havaitsemme, että menetelmän universaalisuus perustuu Kolmogorov-kompleksisuuden ratkeamattomuuteen, joten käytännön sovellusten tulee perustua eri approksimointimenetelmiin.

Kolmannessa luvussa käy ilmi, että mikä tahansa pakkausalgoritmi approksimoi Kolmogorov-kompleksisuutta. Luvussa määritellään normalisoitu kompressioetäisyys kahden objektin välisen etäisyyden mittaamiseen, jossa voidaan käyttää hyväksi mitä tahansa pakkausalgoritmia. Menetelmä on vain näennäisuniversaali, mutta kuitenkin käyttökelpoinen monenlaisen datan vertailuun. Tulosten luokitteluun esitetään erityinen ryvästysheuristiikka, Kvartetti-menetelmä.

Lopuksi yhteenvedossa esitellään mihin menetelmää on jo sovellettu sekä pohditaan sen universaalin luonteen yleispätevyyttä.

2 Algoritminen informaatioteoria

2.1 Kolmogorov-kompleksisuus

Merkkijonon x algoritminen kompleksisuus eli *Kolmogorov-kompleksisuus* määritellään lyhimmän ohjelman p pituutena, joka muodostaa syötteenä saadun merkkijonon x. Mitta siis kuvaa tarkasteltavalle objektille pituudeltaan lyhimmän mahdollisen nollista ja ykkösistä koostuvan merkkijonoesityksen, jonka sisältämä informaatio itsessään riittää muodostamaan alkuperäisen objektin.

Mittaa kutsutaan universaaliksi, koska mitattavan objektin muodostamiseen tarvittava informaatio sellaisenaan kuvailee objektin täydellisesti. Mitta on teoreettinen ja käytännössä mahdoton ratkaista, sillä sen laskeminen on ratkeamaton ongelma. Käytännössä mikä tahansa tietoa pakkaava ohjelma (zip,rar) approksimoi pakattavan kohteen Kolmogorov-kompleksisuutta.

Mitan universaalisuus perustuu merkkijonojen algoritmisiin ominaisuuksiin. Kukin kuvattava objekti voidaan ajatella merkkijonona, joka voidaan esittää nollien ja ykkösten eli bittien sarjana. Kaikki mahdolliset bittisarjat voidaan kuvata vastaamaan luonnollisia lukuja.

Palautetaan mieliin universaali Turingin kone ja rekursiiviset sekä osittaisrekursiiviset funktiot. Funktio on rekursiivinen, jos funktion arvon laskemiseksi on olemassa Turingin kone, joka pysähtyy kaikilla mahdollisilla syötteillä. Funktio on osittaisrekursiivinen, jos sen laskenta pysähtyy jollakin mahdollisella syötteellä. Universaali Turingin kone simuloi mitä tahansa Turingin konetta. Se saa syötteenään simuloitavan Turingin koneen koodin sekä syötteen. Universaali Turingin kone on konstruoitavissa tehokkaasti [Orp94, s.84].

Olkoon S merkkijonojoukko ja numeroidaan sen alkiot $x_i \in S$ yksikäsitteisesti jonkin luonnollisille luvuille kuvautuvan menetelmän $n(x)$ avulla. Kuvauksien

vertailua varten määritellään osittaisfunktio f, joka saa syötteenään luonnollisen luvun p ja joka tulostaa vertailuluvun n, $n = f(p)$. Luvun p pituutta merkitään $l(p)$.

Joukon S merkkijonon x kompleksisuudelle C pätee kuvausfunktion f nojalla:

$$C_f(x) = min\{l(p) : f(p) = n(x)\},$$

missä $C_f(x) = \infty$, jos sopivaa lukua p ei ole olemassa. Tietojenkäsittelyn termein p voidaan ajatella ohjelmana ja f tietokoneena siten, että $C_f(x)$ on lyhimmän ohjelman p pituus, jonka suoritus tietokoneella f tulostaa x:n.

Kuvauksien $f_1, f_2, ... , f_n$ vertailua varten voidaan konstruoida menetelmä f, jonka kuvaama kompleksisuus $C(x)$ kaikille x on vain vakiotermin c verran suurempi kuin pienin mahdollinen kompleksisuus kaikkien kuvausmenetelmien f_i joukosta. Kuvausmenetelmä f *minimoi (minorizes)* menetelmän g, jos kaikille x on olemassa vakio c, siten että $C_f(x) \leq C_g(x) + c$. Kuvausmenetelmät ovat ekvivalentteja, jos ne minimoivat toinen toisensa [LiV97, s. 94].

Universaalin kompleksisuusmitan kannalta on olennaista, että kuvausmenetelmien ekvivalenssiluokkien hierarkiassa on luokka, joka minimoi kaikki muut kuvausmenetelmät. Tällöin hierarkiassa on yksikäsitteinen minimaalinen ominaisuus.

Olkoon $\mathcal{C}$ jokin joukko sellaisia osittaisfunktioita, jotka kuvautuvat positiivisille kokonaisluvuille. Funktio u on universaali joukossa $\mathcal{C}$, jos $u \in \mathcal{C}$ ja jos kaikille $g \in \mathcal{C}$ on olemassa vakio $c_{u,g}$ siten, että $C_u(x) \leq C_g(x) + c_{u,g}$ kaikille x. Vakio $c_{u,g}$ riippuu siis vain funktioista u ja g, ei syötteestä x. Tällöin kahdelle joukon S universaalille funktiolle u ja g pätee $|C_u(x) - C_g(x)| \leq c_{u,g}$. Nyt kompleksisuuden $C(x)$ määrittämiseen voidaan käyttää mitä tahansa universaalia funktiota, sillä minkä tahansa kahden universaalin funktion kuvaama kompleksisuus eroaa vain vakiotermin verran.

Kaikille kuvausmenetelmäjoukoille, kuten positiivisia kokonaislukuja kuvaaville osittaisfunktioiden joukolle, tämä ei kuitenkaan päde. Osoittautuu, että osittaisrekursiivisten funktioiden luokalle ominaisuus pätee. Olkoon ϕ_n jokin osittaisrekursiivinen funktio ja olkoon ϕ_0 universaalin Turingin koneen U laskema funktio.

U saa syötteensä muodossa $< n, p >= 11...10np$, missä ykkösiä on $l(n)$ kertaa. Ohjelma $< n, p >$ on kaksiosainen koodi, jonka alkuosa sisältää Turingin koneen T_n, ja jonka jälkimmäisenä osana on ohjelma p. Näin ollen Universaali kone U voi ensin jaotella syötteensä kahteen osaan, jonka jälkeen se laskea suoritukseen koneen T_n syötteenä p. Toisin sanoen $\phi_0(< n, p >) = \phi_n(p)$. Tällöin on olemassa universaali osittaisrekursiivinen funktio ϕ_0, jos T_n:n suoritus pysähtyy osittaisrekursiiviselle funktiolle ϕ_n, jolloin pätee (universaalisuusteoreema)

$$C_{\phi_0}(x) \leq C_{\phi_n}(x) + c_{\phi_n},$$

missä vakiolle c_{ϕ_n} voidaan asettaa arvo $2l(n) + 1$ [LiV97, s.97].

Lauseen suorana seurauksena saadaan invarianssiteoreema, jonka mukaan kahdelle universaalille kuvausmenetelmälle u ja v pätee

$$|C_u(x) - C_v(x)| \leq c_{u,v}.$$

Universaali kuvausmenetelmä ei siis välttämättä anna lyhintä kuvausta merkkijonolle x, mutta sen intuitiivinen merkitys on siinä, ettei mikään muu kuvaus voi suoriutua paremmin äärettömälle määrälle merkkijonoja. Universaalifunktion antama kompleksisuus eroaa vain vakiotermin verran optimaalisesta kuvauksesta.

Alkuosakoodi (prefix-code) on yksikäsitteisesti tunnistettavissa oleva merkkijonojoukon alkio siten, ettei se ole minkään muun joukon alkion alkuosa. Joukko on *alkuosavapaa (prefix free)* silloin kun joukon mikään alkio ei ole toisensa alkuosa. Esimerkiksi Joukossa $\{0, 01, 11, 101\}$ alkio 0 on alkion 01 alkuosa, mutta joukossa

$\{0, 10, 110, 111\}$ mikään alkio ei ole toistensa alkuosa. Mikä tahansa alkuosakoodi voidaan uudelleenkoodata yksikäsitteisesti [LiV97, s.73].

Kraftin epäyhtälön mukaan jonolle luonnollisia lukuja $l_1, l_2, ...l_n$ on olemassa alkuosakoodit, joiden pituudet vastaavat koodisanojen pituuksia, jos ja vain jos pätee $\sum_n 2^{-l_n} \leq 1$ [LiV97, s.74].

Algoritmisen kompleksisuuden kannalta kuvausmenetelmät on syytä rajoittaa siten, että ne täyttävät alkuosaehdon. Osittaisrekursiivinen alkuosafunktio $\phi : \{0,1\}^* \to \mathcal{N}$ on osittaisrekursiivinen funktio siten, että jos $\phi(p) < \infty$ ja $\phi(q) < \infty$, niin p ei ole q:n alkuosa [LiV97, s.192].

Universaalisuus pätee myös osittaisrekursiivisille alkuosafunktioille. Voidaan osoittaa, että on olemassa universaali osittaisrekursiivinen alkuosafunktio ψ_0 siten, että mille tahansa osittaisrekursiiviselle alkuosafunktiolle ψ on olemassa vakio c_ψ, jolloin

$$C_{\psi_0}(x|y) \leq C_\psi(x|y) + c_\psi$$

pätee kaikille $x, y \in \mathcal{N}$ [LiV97, s.193].

Nyt voidaan määritellä ehdollinen invarianssiteoreema, joka on soveltamisen kannalta hyödyllinen tulos. Kahdelle universaalille osittaisrekursiivisille alkuosafunktioille pätee

$$|C_{\psi_1}(x|y) - C_{\psi_2}(x|y)| \leq c_{\psi_1,\psi_2}.$$

Invarianssiteoreeman ehdollisen määritelmän intuitiivinen merkitys on siinä, että jos osittaisrekursiiviset (alkuosa)funktiot ajatellaan ohjelmointikielinä, niin minkä tahansa kahden valitun ohjelmointikielen välillä tarkasteltavan ohjelman kompleksisuus eroaa korkeintaan vakiotermin verran. Näin ollen ohjelman kompleksisuus ei riipu ohjelmointikielen valinnasta, vaan mitä tahansa kieltä voidaan käyttää.

Invarianssiteoreeman nojalla Kolmogorov-kompleksisuus on referenssifunktion

valintaa vaille optimaalinen kompleksisuuden kuvausmenetelmä kaikille merkkijonoille x, missä referenssifunktion valinta aiheuttaa korkeintaan vain vakiotermisen menetyksen optimaaliseen kuvausmenetelmään verrattuna. Ehdollinen alkuosa-Kolmogorov-kompleksisuus $K(x)$ määritellään valitsemalla jokin universaali osittaisrekursiivinen alkuosafunktio ψ_u referenssi-koneeksi, jolloin pätee

$$K(x|y) = C_{\psi_u}(x|y),$$

kaikille x [LiV97, s.194].

Osittaisrekursiivisten funktioiden rajoittaminen alkuosaehdolla tekee kompleksisuuden tarkastelun tietyissä tilanteissa realistisemmaksi ja johdettavien tulosten kannalta suotuisammaksi. Esimerkiksi lisättävää vakiotermiä suuruusluokaltaan $O(1)$ vaille pätee $K(x, y) \leq K(x) + K(y)$, kun osittaisrekursiivisille funktioille suuruusluokka olisi logaritminen [LiV97, s.194].

Lisäksi voidaan osoitaa, että suuruusluokaltaan $O(log(K(y)))$ lisättävää vakiotermiä vaille pätee $K(x|y) = K(x, y) - K(y)$ [Ben98, s.1410].

2.2 Kelvollinen etäisyysmitta

Ehdollinen Kolmogorov-kompleksisuus $K(x|y)$ ilmaisee lyhimmän ohjelman pituuden, joka muodostaa x:n annettuna y. Pituus ei kuitenkaan sellaisenaan riitä etäisyysmitaksi kahden satunnaisen objektin välille, sillä $K(y|x)$ ei välttämättä ole yhtä suuri kuin $K(x|y)$ ja ei näin ole symmetrinen.

Seuraavassa aliluvussa kuvataan normalisoitu informaatioetäisyys, joka tarjoaa teoreettisen kehyksen kahden objektin välisen etäisyyden tarkasteluun. Seuraava luku esittelee edellisen approksimaationa normalisoidun kompressioetäisyyden, jonka avulla voidaan kahden objektin välistä etäisyyttä mitata käytännössä, jolloin objektijoukon keskinäiset etäisyydet voidaan luokitella esimerkiksi etäisyys-

matriisiin. Sitä ennen määritellään ehdot, jotka algoritmisesti mielekkään etäisyysmitan ja mitta-avaruuden täytyy toteuttaa.

Olkoon $\Omega = \{0,1\}^*$. Funktio $D : \Omega \times \Omega \rightarrow R^+$ on kuvaus positiivisille reaaliluvuille. Määritellään ehdot algoritmisessa mielessä *kelvolliselle (admissible)* etäisyysmitalle. Etäisyys D on kelvollinen mitta, jos kaikille $x, y, z \in \Omega$ pätee [CiV05, s.1525]:

i) $D(x,y) = 0 \equiv x = y$ (identiteettiaksiooma),

ii) $D(x,y) = D(y,x)$ (symmetria-aksiooma).

iii) $d(x,y) + d(y,z) \geq d(x,y)$ (kolmioepäyhtälö),

iv) D on semi-laskettavissa ylärajansa suhteen (upper semi-computable),

v) kullekin objektiparille $x, y \in \Omega$ etäisyys $D(x,y)$ on pituus alkuosakoodille, jonka kuvaama ohjelma muodostaa x:n avulla y:n sekä y:n avulla x:n, valitun referenssifunktion (ohjelmointikielen) nojalla.

Etäisyyden normalisointi asettaa lisävaatimuksen. Olkoon funktio d normalisoitu etäisyysmitta, joka kuvaa karteesiselle tulolle arvon väliltä $\Omega \times \Omega \rightarrow [0,1]$. Tällöin etäisyys d on kelvollinen normalisoitu etäisyysmitta, jos se toteuttaa kohdat $i - v$ ja lisäksi normalisointiehdon:

vi) d on normalisoitavissa siten, että yläraja etäisyyksien summalle on 1. Tällöin kaikille vakioille $e \in [0,1]$ pätee $\mid \{y : d(x,y) \leq e \leq 1\} \mid < 2^{eK(x)+1}$ [CiV05, s.1526].

2.3 Normalisoitu informaatioetäisyys

Määritellään normalisoitu informaatioetäisyys, joka toteuttaa kelvollisen mitta-avaruuden ehdot $i - vi$.

Funktio $I_K(x:y) = K(y) - K(y|x)$ kuvaa y:n algoritmista informaatiota, joka sisältyy x:ään [LiV97, s.233]. Osoittautuu, että lisättävää logaritmista vakiotermiä vailla pätee $K(x,y) = K(x)+K(y|x) = K(y)+K(x|y)$, josta seuraa, että lisättävää logaritmista virhearvoa vaille pätee $K(x)-K(x|y) = K(y)-K(y|x)$. Sanotaan, että y:llä ja x:llä on toisistaan lähes yhtä paljon informaatiota. Tätä kutsutaan niiden väliseksi *keskinäisinformaatioksi (mutual information)*. Keskinäisinformaatiota on sitä vähemmän mitä lähempänä arvo on nollaa, jolloin merkkijonot ovat toisistaan riippumattomia. [Ben98, s.1410].

Olkoon $E(x,y)$ kahden merkkijonon välinen algoritminen informaatioetäisyys, joka kuvaa lyhimmän binäärisen ohjelman, joka tulostaa x:n syötteenään y ja toisin päin. Ohjelma siis pitää sisällään kaiken tarvittavan tiedon siitä, miten syötteestä x muodostuu y ja syötteestä y muodostuu x, kuitenkaan kykenemättä muuttumaan mitenkään suorituksen aikana [Ben98, s.1410].

$E(x,y)$ on teoreettisessa mielessä semi-laskettavissa ylärajansa suhteen (kohta iv), sillä voidaan löytää yhä lyhyempi alkuosakoodi ohjelmalle p, joka laskee $p(x) = y$ ja $p(y) = x$, kun aikavaativuutta ei ajatella esteenä kaikkien mahdollisten ohjelmien suorittamisessa [Ben98, s.1415].

Olkoon E kelvollinen etäisyys. Tällöin kaikille $x \in \{0,1\}^*$ joukko $\{E(x,y) : y \in \{0,1\}^*\}$ on alkuosakoodien pituuksien joukko, jolloin se toteuttaa ehdon v Kraftin epäyhtälön nojalla $\sum_y 2^{-E(x,y)} \leq 1$ [CiV05, s.1526].

Suuruusluokaltaan logaritmisen vakiotermin $O(log \max\{K(y|x), K(x|y)\})$ säteellä pätee [Ben98, s.1408]

$$E(x,y) = \max\{K(y|x), K(x|y)\}.$$

Määritellään normalisoitu informaatioetäisyys NID seuraavasti:

$$NID(x,y) = \frac{E(x,y)}{\max\{K(x), K(y)\}} = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}.$$

Jos $K(y) \geq K(x)$, niin $NID(x,y) = \frac{K(y)-I(x:y)}{K(y)} = 1 - \frac{I(x:y)}{K(y)}$, joka ilmaisee x:n ja y:n jakaman keskinäisinformaation suhteessa kompleksisuudelta suurempaan merkkijonoon [LCL+04, s.3254].

Normalisoidulle informaatioetäisyydelle pätee kelvollisen mitta-avaruuden ehdot. Symmetrisyys on triviaali. Identiteettiaksiooma $NID(x,x) = 0$ toteutuu suuruusluokaltaan $O(\frac{1}{K(x)})$ tarkkuudella. Kolmioepäyhtälön monimutkainen todistus on esitetty lähteessä [LCL+04, s.3254], jonka mukaan yhtälö pätee suuruusluokaltaan korkeintaan $O(\frac{1}{\max\{K(x),K(y),K(z)\}})$ lisättävän vakiotermin tarkkuudella.

Selvästikin etäisyys d saa arvot väliltä $[0, 1 + O(\frac{1}{\max\{K(x),K(y)\}})]$. On osoitettu, että normalisointiehto (vi) pätee ja näin ollen d on normalisoitavissa [LCL+04, 3255].

NID on universaali, koska sen voidaan osoittaa minimoivan kaikkia ylhäältä semilaskettavissa olevia normalisoituja etäisyyksiä $f(x,y)$ siten, että

$$d(x,y) \leq f(x,y) + c,$$

missä $c = \frac{1}{\min K(x),K(y)}$ [LCL+04, s.3256].

3 Ryvästys kompressoinnin avulla

3.1 Normalisoitu kompressioetäisyys

Normalisoitu informaatioetäisyys osoittautuu algoritmisessa mielessä ratkeamattomaksi ongelmaksi, mutta sitä voidaan approksimoida. Merkkijonon Kolmogorov-kompleksisuus on teoreettinen raja sille kuinka pieneksi mikä tahansa kompressointialgoritmi voi merkkijonon tiivistää. Näin ollen Kolmogorov-kompleksisuutta voidaan approksimoida eri pakkausmenetelmin.

Normaalikompressori C on pakkausmenetelmä, joka toteuttaa seuraavat ehdot.

$C(xx) = C(x)$ ja $C(\lambda) = 0$, missä λ on tyhjä merkkijono. Monotonisuus: $C(xy) > C(x)$. Symmetria $C(xy) = C(yx)$. Distributiivisuus: $C(xy)+C(z) \leq C(xz)+C(yz)$ [CiV05, s.1527].

Arvo $C(y|x)$ kuvaa merkkijonon y informaatiota suhteessa merkkijonoon x ja sitä voidaan approksimoida kaavalla $C(y|x) = C(xy) - C(x)$ [CiV05, s.1527].

Pakkausmenetelmiä käsiteltäessä voidaan merkkijonot x ja y katenoida, sillä lisättävää logaritmista vakiotermiä vaille pätee $K(x, y) = K(xy) = K(yx)$.

Normalisoidun informaatioetäisyyden kaavan osoittaja $\max\{K(x|y), K(y|x)\}$ voidaan kirjoittaa muotoon $\max\{K(x, y) - K(x), K(x, y) - K(y)\}$, joka pätee lisättävän logaritmisen vakiotermin säteellä. On esitetty, että kaava on käytännössä parhaiten approksimoitavissa kaavalla $\min\{C(xy), C(yx)\} - \min\{C(x), C(y)\}$. Lisäksi jaksokoodaus-kompressorit (block-coding) ovat melkein symmetrisiä jo määritelmänsä mukaan, jolloin etäisyyttä $\min\{C(xy), C(yx)\}$ voidaan approksimoida pelkällä $C(xy)$ tai $C(yx)$ arvolla [CiV05, s.1528].

Määritellään kompressioetäisyys $E_C(x, y)$, joka approksimoi informaatioetäisyyttä $E(x, y)$, referenssikompressorin C nojalla seuraavasti:

$$E_C(x, y) = C(xy) - \min\{C(x), C(y)\},$$

missä $C(xy)$, $C(x)$ ja $C(y)$ kuvaavat kompressoitujen merkkijonojen xy, x ja y pituutta. Voidaan todistaa, että $E_C(x, y)$ on kelvollinen etäisyys lisättävää suuruusluokaltaan $O(1)$ vakiotermin säteellä, jos C täyttää normaalikompressorin ehdot [CiV05, s.1528].

Sovelletaan Normalisoidun informaatioetäisyyden kaavaa ja approksimoidaan Kolmogorov-kompleksisuutta normaalikompressorilla C. Saadaan normalisoitu kompressioetäisyys NCD:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

Voidaan todistaa, että NCD on kelvollinen normalisoitu etäisyys silloin, kun käytetään normaalikompressorin ehdot täyttävää pakkausmenetelmää [CiV05, s.1529].

Käytännössä NCD antaa kahden objektin vertailuun arvon väliltä $0 \leq r \leq 1 + \epsilon$: mitä pienempi arvo on, sen enemmän vertailtavissa esiintyy samankaltaisuutta. Ylärajan vakio ϵ johtuu pakkaustekniikoiden vajaavaisuudesta. Perinteisten pakkausalgoritmien suhteen vakio ϵ ylittää erittäin harvoin arvon 0.1 ja esimerkiksi PPMZ pakkaaja on käytännön kokeissa antanut enintään arvon 1 [CiV05, s.1529].

NCD on näennäisuniversaali siinä mielessä, että se on NID:n approksimaatio. Objektin Kolmogorov-kompleksisuutta ei voida ratkaista, ja näin ollen NCD:n tuloksia ei voida verrata teoreettiseen optimiin. Menetelmän tehoa on vaikea arvioida selkeästi muuten kuin käytännön kokein.

NCD:n avulla voidaan vertailtaville objekteille muodostaa esimerkiksi etäisyysmatriisi, jonka alkioissa esitetään kunkin parin etäisyys toisistaan. Etäisyysmatriisista sellaisenaan on vaikea hahmottaa vertailtuja tuloksia, sillä arvot ovat yleensä hyvin lähellä toisiaan esimerkiksi välillä $[0.85, 1.1]$. Datan visualisointiin tarvitaan jokin erityinen menetelmä, joka jatkokäsittelee datan mielekkäästi. Seuraavassa esitetään eräs vaihtoehto.

3.2 Kvartetti-menetelmä

Ryvästys (clustering) on luonteva tapa luokitella samankaltaiset objektit omiin ryppäisiinsä. [CiV05, s.1530] esittää NCD-etäisyysmatriisille räätälöidyn hierarkisen ryvästyksen, jonka heuristiikka perustuu *mäkikiipeilyyn (hill-climbing)* ja joka käyttää satunnaisesti mutaatio-operaatioita approksimointiin. Menetelmää kutsutaan *Kvartetti-menetelmäksi (The quartet method)*.

Menetelmä muodostaa datasta *dendrogrammin (dendrogram)*, joka on suunnattu binääripuu tai suuntaamaton *kolmihaarapuu (ternary tree)*. Sen etuna on, ettei luo-

kiteltavien rypäiden määrää tarvitse tietää etukäteen. Menetelmän ongelmana on vertailtavien alkioiden mahdollinen suuri määrä. Hierarkinen ryvästys toimii vielä siedettävästi noin 25 objektille, mutta jo yli 40 vertailtavaa voi viedä laskennallisesti aivan liian kauan aikaa. Ratkaisuksi ehdotetaan, että ennen kvartettimenetelmän käyttöä datan voisi ryvästää ei-hierarkisesti esimerkiksi *moniulotteisen k-means skaalauksen (multidimensional scaling of k-means)* avulla.

Kvartetti-menetelmä käsittelee vertailtavia alkioita neljän ryhmissä, jolloin n alkioiselle joukolle on olemassa $\binom{n}{4}$ ryhmää. Kullekin ryhmälle u, v, w, x muodostetaan suuntaamaton kolmihaarainen puu, jonka puurakenteessa on kaksi kaksilehtistä alipuuta. Kutsutaan rakennetta kvartettitopologiaksi (quartet topology). Rakenne on mahdollista muodostaa kolmella eri tavalla:

i) $uv|wx$ ii) $uw|vx$ iii) $ux|vw$.

Mikä tahansa puu T on konsistentti, jos mille tahansa neljän alkion ryhmän u, v, w, x alipuulle $uv|wx$ pätee, että polku u:sta v:hen ei ylitä polkua w:sta x:ään. Näin ollen mille tahansa neljän alkion ryhmälle tasan yksi kvartettitopologiavaihtoehdoista on konsistentti mille tahansa puulle. Kvartetti-menetelmän tavoite on löytää puu, joka sisältää maksimaalisen määrän konsistentteja kvartettitopologioita.

Määritellään minimaalisen kvartettipuun kustannuksen optimointinongelma (minimum quartet tree cost) MQC seuraavasti. Kvartettitopologian kustannus määritellään kunkin naapuriparin summana, $C_{uv|wx} = d(u, v) + d(w, x)$. Puun T yhteiskustannus C_T N-alkioiselle joukolle N-lehtisolmuja määritellään sen kaikkien konsistenttien kvartettitopologioiden kustannusten summana:

$$C_T = \sum_{\{u,v,w,x\} \subseteq N} \{C_{uv|wx} : \text{ T on konsistentti, kun } uv \mid wx\}.$$

Luodaan lista kaikista mahdollisista kvartettitopologioista ja lasketaan kunkin neljän alkion ryhmän kolmelle mahdolliselle kvartettitopologialle paras (minimaalinen) kustannus

$$m(u, v, w, x) = min\{C_{uv|wx}, C_{uw|vx}, C_{ux|vw}\}$$

ja huonoin (maksimaalinen) kustannus

$$M(u, v, w, x) = max\{C_{uv|wx}, C_{uw|vx}, C_{ux|vw}\}.$$

Summaamalla kaikki parhaat kvartettitopologiat saadaan paras (minimaalinen) kustannus

$$m = \sum_{\{u,v,w,x\} \subseteq N} m(u, v, w, x)$$

ja päinvastoin summaamalla kaikki huonoimmat kvartettitopologiat saadaan huonoin (maksimaalinen) kustannus

$$M = \sum_{\{u,v,w,x\} \subseteq N} M(u, v, w, x).$$

Vertailun helpottamiseksi skaalataan kustannus lineaarisesti siten, että huonoin kuvautuu arvoon nolla ja paras arvoon 1, ja merkitään normalisoitua arvoa $S(T) = \frac{M-C_T}{M-m}$. Tavoitteena on löytää täysi puu, jolla on maksimaalinen arvo $S(T)$, joka on siis matalin yhteiskustannus. Osoittautuu, että ongelma on laskennallisesti NP-kova, mutta se on joskus ratkaistavissa ja aina approksimoitavissa esimerkiksi seuraavan heuristiikan avulla.

Heuristiikka perustuu puurakenteeseen kohdistuviin satunnaisesti suoritettaviin mutaatio-operaatiohin ja mäkikiipeilyyn. Ensin luodaan satunnaisesti puu T, jossa on $2n - 2$ solmua ja joka sisältää n lehtisolmua, jotka nimetään data-alkioiden mukaan. Puussa on $n - 2$ sisäsolmua, joita merkitään n. Jokaisella sisäsolmulla on tasan kolme yhdistävää polkua.

Puun T kvartettitopologioille lasketaan kustannukset, jonka jälkeen edellämainituin keinoin lasketaan $S(T)$. Määritellään puumutaatio, joka voi olla jokin seuraavista muutoksista:

Lehden vaihto, jossa kaksi satunnaisesti valittua lehteä vaihdetaan keskenään.

Alipuun vaihto, jossa kaksi satunnaisesti valittua sisäsolmua ja niiden alipuut vaihdetaan keskenään.

Alipuun siirto, jossa satunnaisesti valittu alipuu (voi olla lehti) irroitetaan ja liitetään toiseen paikkaan siten, että rakenne-ehdot säilyvät muuttumattomina.

Kukin mutaatio pitää lehtisolmujen ja sisäsolmujen määrät muuttumattomina, jolloin vain rakenne ja sijainnit vaihtuvat.

Suoritetaan puulle T täysimutaatio, joka koostuu sarjasta puumutaatioita, jotka valitaan seuraavan jakauman mukaan. Ensin valitaan satunnaisesti luku k, jonka mukaan valitaan k puumutaatiota. Puumutaatioiden muodostamaa jaksoa suoritetaan todennäköisyydellä 2^{-k}. Kullekin mutaatiolle valitaan kolmesta mutaatiotyypistä yksi samalla todennäköisyydellä $\frac{1}{3}$. Lopuksi kukin mutaatio suoritetaan yhdenmukaisesti ja satunnaisvalinnalla arvotaan mahdollisten uusien lehtisolmujen tai sisäsolmujen paikat siten, että puurakenteen ehdot säilyvät.

Nyt saadaan puu $T^{'}$, jolle lasketaan $S(T^{'})$, ja jos $S(T^{'}) > S(T)$, niin pidetään $T^{'}$ parhaana puuna ja jatketaan iteratiivisesti mutaatioiden suorittamista, kunnes S(T) saavuttaa maksimiarvon yksi, jolloin konsistentti puurakenne on löytynyt. Lisäksi approksimaatiolle määritellään aikaraja- ja (tai) kustannusvaatimukset, joiden täytyttyä lopetetaan iteraatio.

Menetelmä antaa yleensä puolen tunnin sisällä tyypillisille korkeintaan 40-alkioisille tosimaailman ongelmille puun, jonka $S(T) > 0.9$.

4 Yhteenveto

Seminaarityö esitteli datan kompressointiin perustuvan, universaaliksikin kutsutun menetelmän kahden objektin samankaltaisuuden tarkasteluun sekä Kvartettimenetelmän samankaltaisen tiedon ryvästämiseen. Menetelmää voidaan tavallisuudesta poiketen soveltaa sellaisenaan useisiin tiedon louhinnan ongelmiin.

Tiedon louhinnan algoritmeissa käytetään paljon hyväksi ennalta-asetettuja parametreja. Nämä subjektiiviset päätökset voivat aiheuttaa etenkin seuraavanlaisia ongelmia. Ensinnäkin parametrien väärät arvot voivat johtaa algoritmeja muodostamaan vääristyneitä malleja. Toiseksi algoritmit voivat muodostaa malleja, joita ei todellisuudessa esiinny tai päätyvät yliarvioimaan löydetyn mallin merkityksellisyyttä. Lisäksi voi olla erittäin vaikeaa verrata eri algoritmien samasta datasta muodostamia tuloksia keskenään [KLR04, s.1].

Kompressiontiin perustuva lähestymistapa pyrkii ratkaisemaan näitä ongelmia. Menetelmä nojautuu vahvasti algoritmiseen informaatioteoriaan ja sen tulosten approksimointiin. Sen universaalisuus perustuu datan pakkautuvuuteen, joka on kaikelle bitteinä esitettävissä olevalle tiedolle yhteinen ominaisuus. Menetelmästä tekee näennäisuniversaalin se, että käytäntö perustuu vain teoreettisen optimin approksimointiin ja sen hyvyyttä on vaikea arvioida. Kuitenkin moni perinteinen pakkausalgoritmi on riittävän hyvä approksimoimaan teoreettista optimia ja käytännön tulokset eivät kaadu pakkaajan vajaavuuteen.

Periaatteessa menetelmää voidaan soveltaa missä tahansa tiedon louhinnan ympäristössä. Käytännössä ongelmia aiheuttaa menetelmän raskaus suurelle verrattavien objektien joukolle. Menetelmästä voi tulla nopeasti liian hidas.

Menetelmä on kuitenkin luonteeltaan ainutlaatuinen ja tiedon louhijan on hyvä tietää sen olemassaolosta. Sen avulla voi pyrkiä saamaan hyviä aputuloksia tai mahdollisesti suuntaa-antavia vertailutuloksia perinteisiin menetelmiin nähden.

Menetelmä on kehitetty nykyiseen muotoonsa vasta 2000-luvun alussa ja sitä on sovellettu vielä hyvin vähän. Sitä on sovellettu muun muassa kielitieteissä ja bioinformatiikassa. Sen avulla on voitu tunnistaa satunnaisessa tekstiotoksessa käytetty kieli, analysoida perimän vaikutuksia ihmisessä ja jopa tutkia sairauksia aiheuttavien virusten alkuperää. Erinäisiä musiikkikappaleita on voitu tunnistaa säveltäjänsä mukaan. Lisäksi menetelmä on poikinut määritelmän normalisoidulle Google-etäisyydelle, jossa Internetin Google-hakua ja informaatioetäisyyttä käytetään hyväksi etsimään haetuille sanoille merkitystä.

Lähteet

Ben98 Bennett, C., et. al., Information distance. *IEEE Transactions on information theory, 44, 4(1998), 1407-1423.*

CiV05 *Cilibrasi, R., Vitányi, P., Clustering by compression.* IEEE Transactions on Information Theory, 51, 4(2005), sivut 1523-1545.

KLR04 Keogh, E., Lonardi, S., Rtanamahatana, C.A., Towards parameter-free data mining. *Proc. 10th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining, Seattle, Washington, USA, elokuu 2004, sivut 206-215.*

LCL+04 *Li, M., Chen, X., Li, X., Ma, B., Vitányi, P., The Similarity metric.* IEEE Transactions on Information Theory, 50, 12(2004), sivut 3250-3264.

LiV97 Li, M., Vitányi, P., *An Introduction to Kolmogorov Complexity and Its Applications, 2nd Edition. Springer-Verlag, New York, 1997.*

Orp94 *Orponen, P.,* Laskennan teoria. Helsingin yliopisto, Tietojenkäsittelytieteen laitos, Helsinki, 1994.