Normalisoitu pakkausetäisyys

Timo Sand

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 22. syyskuuta 2013

Kolmogorov kompleksisuus Lyhimmän binääriohjelman pituus, joka palauttaa x syötteellä y, on *Kolmogorov kompleksisuus* x:stä syötteellä y; tämä merkitään $K(x|y)$. Pohjimmillaan Kolmogorov kompleksisuus tiedostosta on sen äärimmäisesti pakatun version pituus.

Normalisoitu Informaatioetäisyys Artikkelissa [CV05] on esitelty *informaatioetäisyys* $E(x, y)$, joka on määritelty lyhimpänä binääriohjelmana, joka syötteellä x laskee y:n ja syötteellä y laskee x:n. Tämä lasketaan seuraavasti

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

Normalisoitu versio informaatioetäisyydestä ($E(x, y)$), jota kutsutaan *normalisoiduksi informaatioetäisyydeksi*, on määritelty seuraavasti

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

Tätä kutsutaan *samankaltaisuuden metriikaksi*, koska tämän on osoitettu [CV05] täyttävän vaatimukset etäisyyden metriikaksi. *NID* ei kuitenkaan ole laskettavissa tai edes semi-laskettavissa, koska Turingin määritelmän mukaan Kolmogorov kompleksisuus ei ole laskettavissa. Nimittäjän approksimointi annetulla kompressorilla C on $max\{C(x), C(y)\}$. Osoittajan paras approksimaatio on $max\{C(xy), C(yx)\} - \min\{C(x), C(y)\}$ [CV05]. Kun *NID* approksimoidaan oikealla kompressorilla, saadaa tulos jota kutsutaan *normalisoiduksi pakkausetäisyydeksi*. Tämä esitellään formaalisti myöhemmin.

Normaali Kompressori Seuraavaksi esitämme aksioomia, jotka määrittelevät laajan joukon kompressoreita ja samalla varmistavat *normalisoidussa pakkausetäisyydessä* halutut ominaisuudet. Näihin kompressoreihin kuuluvat monet tosielämän kompressorit.

Kompressori C on *normaali* jos se täyttää seuraavat aksioomat, $O(logn)$ termiin saakka:

1. *Idempotenssi*: $C(xx) == C(x)$ ja $C(\lambda) = 0$, jossa λ on tyhjä merkkijono,

2. *Monotonisuus*: $C(xy) \geq C(x)$,

3. *Symmetrisuus*: $C(xy) == C(yx)$ ja

4. *Distributiivisuus*: $C(xy) + C(z) \leq C(xz) + C(yz)$.

Normalisoitu Pakkausetäisyys Normalisoitua versiota *hyväksyttävästä etäisyydestä* $E_c(x, y)$, joka on kompressoriin C pohjautuva approksimaatio *normalisoidusta informaatioetäisyydestä*, kutsutaan nimellä *Normalisoitu Pakkausetäisyys (NCD)* [CV05]. Tämä lasketaan seuraavasti

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

NCD on funktioden joukko, joka ottaa argumenteiksi kaksi objektia (esim. tiedostoja tai Googlen hakusanoja) ja tiivistää nämä, erillisinä ja yhdistettyinä. Tämä funktioden joukko on parametrisoitu käytetyn kompressorin C mukaan.

Käytännössä *NCD*:n tulos on välillä $0 \leq r \leq 1 + \epsilon$, joka vastaa kahden tiedoston eroa toisistaan; mitä pienempi luku, sitä enemmän tiedostot ovat samankaltaisia. Tosielämässä pakkausalgoritmit eivät ole yhtä tehokkaita kuin teoreettiset mallit, joten virhemarginaali ϵ on lisätty ylärajaan. Suurimmalle osalle näistä algoritmeistä on epätodennäköistä että $\epsilon > 0.1$.

Luonnollinen tulkinta *NCD*:stä, jos oletetaan $C(y) \geq C(x)$, on

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

Eli etäisyys x:n ja y:n välillä on suhde y:n parannuksesta kun y pakataan käyttäen x:ää, ja y:n pakkauksesta yksinään; suhde ilmaistaan etäisyytenä bittien lukumääränä kummankin pakatun version välillä.

Kun kompressori on normaali niin *NCD* on normalisoitu hyväksyttävä etäisyys, joka täyttää metriikan yhtälöt, eli se on samankaltaisuuden metriikka.

Lähteet

[CV05] Cilibrasi, Rudi ja Vitanyi, Paul M. B.: *Clustering by Compression.* IEEE Transactions on Information Theory, 51(4):1523–1545, Huhtikuu 2005.