Normalisoitu pakkausetäisyys

Timo Sand

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 21. 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.

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

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

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

NID:iä 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 Kolmogorv kompleksisuus ei ole laskettavissa. Nimittäjän approksimointi annetulla kompressorilla C on trviaalia, se on $max\{C(x), C(y)\}$. Numeraattori on hankalampi ja sen 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 Esitämme aksioomia jotka määrittelevät laajan joukon kompressoreita joihin kuuluvat moni tosielämän kompressori ja samalla varmistavat *NCD*:ssä halutut ominaisuudet.

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

1. *Idempotency*: $C(xx) == C(x)$ ja $C(\lambda) = 0$, jossa λ on tyhjä merkkijono.
2. *Monotonisuus*: $C(xy) \geq C(x)$
3. *Symmetrisuus*: $C(xy) == C(yx)$

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

Normalisoitu Pakkausetäisyys Normalisoitua versio *hyväksyttävästä etäisyydestä* $E_c(x, y)$, joka on kompressoriin C pohjautuva approksimaatio *Normalisoidusta Informaatioetäisyydestä (NID)*, kutsutaan nimellä *Normalisoitu Pakkausetäisyys (NCD)* [CV05]

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

NCD on funktioden perhe jotka ottavat argumenteiksi kaksi objektia (esim. tiedostoja, Googlen haku sanoja) ja tiivistävät nämä, erillisinä ja yhdistettyinä. Tämä funktioden perhe on parametrisoitu käytetyn kompressorin mukaan.

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

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 pituutena bitteinä kummankin pakatun version välillä.

Kun kompressori on normaali, sitten *NCD* on normalisoitu hyväksyttävä etäisyys joka täyttää metriikan (epä)yhtälöt, eli 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.