Normalisoitu pakkausetäisyys

Timo Sand

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 20. syyskuuta 2013

Kolmogorov kompleksisuus Lyhimmän binääriohjelman pituus, joka palauttaa x syötteellää 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 informaatioetäisyys 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 Kolmogorv kompleksisuus ei ole laskettavissa Turingin määritelmän mukaan.

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 evaluoivat määrätyn kaavan, joka ilmaisee tiivistetyn version näistä objekteista, erillisinä ja yhdistettynä. Tämä funktioden perhe on parametrisoitu käytetyn kompressorin mukaan.

Käytännössä *NCD*:n tulos on $\leq r \leq 1+\epsilon$ joka kuvastaa kahden tiedoston erotusta; pienempi luku tarkoittaa että tiedostot ovat samankaltaisia. ϵ on virhemarginaali korjaamaan tosielämän pakkausalgoritmejen puutteita, suurimmalle osalle pakkausalgritmeistä on epätodennäköistä nähdä $\epsilon \geq 0.1$

NCD:stä on luonnollinen tulkinta. Jos oletetaan $C(y) \geq C(x)$, niin voimme kirjoittaa

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

Eli siis etäisyys $NCD(x,y)$ x:n ja y:n välillä tuottaa parannuksen kun pakataan y käyttäen x:ää kuten aikaisemmin pakattuna "tietokantana" ja pakkaamalla y tyhjästä, ilmaistuna suhteena pituuden biteissä välillä kummastakin pakatusta versiosta.

Lähteet

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