Referaatti: Clustering by Compression

Timo Sand

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 18. syyskuuta 2013

1 Normalisoitu Pakkausetäisyys

Normalisoitu Pakkausetäisyys (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. Täten tämä funktioden perhe on parametrisoitu käytettävän kompressorin mukaan.

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]

Jos x ja y ovat kaksi objektia, ja $C(x)$ on x kompressoitu etäisyys käyttäen kompressoria C, sitten NCD määräytyy seuraavanlaisesti

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

Oletamme että objektit ovat äärellisiä merkkijonoja 0:sta ja 1:stä. Jokainen tiedosto tietokoneella on tätä muotoa. Voidaan määritellä informaatioetäisyys merkkijonojen x ja y välillä lyhimmän ohjelman p mukaan, joka laskee x:n y:stä ja toisinpäin. Tämä lyhin ohjelma on määrätyssä ohjelmointikielessä. Teknisistä syistä käytetään Turing koneitten teoreettista käsitettä. p:n pituuden ilmaisemiseen käytetään *Kolmogorov kompleksisuuden* käsitettä. On osoitettu, että

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

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

Normalisoitu Informaatioetäisyys

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

,

jossa $K(x|y)$ on algoritmin tietoa x:stä kun y on syöte. *NID*:iä kutsutaan *samankaltaisuuden metriikaksi*, koska $NID(x, y)$ on osoitettu täyttävän vaatimukset etäisyyden metriikaksi. *NID* ei kuitenkaan ole laskettavissa tai edes semi-laskettavissa. *NCD* on siis käytännön versio *NID*:stä

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.