Normalisoitu pakkausetäisyys

Timo Sand

Kandidaatintutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 20. lokakuuta 2013

Sisältö

1 Johdanto

Kaikki data on luotu samanveroiseksi, mutta jotkut ovat samankaltaisempia kuin toiset. Esitämme tavan jolla esittää tämä samankaltaisuus, käyttäen uutta samankaltaisuuden metriikkaa(engl. similarity metric) joka perustuu tiedoston pakkaamiseen. Se on parametriton siten että se ei käytä ominaisuuksia tai taustatieto datasta, ja sitä voi soveltaa eri alueisiin ja aluerajojen yli ilman muunnoksia. Se on yleinen siten että se approksimoi parametrin joka esittää samankaltaisuutta hallitsevassa piirteessä kaikissa pareittain vertailuissa. Se on vakaa siinä mielessä, että sen tulokset ovat riippumattomia käytetystä kompressorista.

Pakkaukseen perustuva samankaltaisuus (engl. Compression-Based Similarity): Tällaisen 'universaalin' metriikan kehittivät [CV05]. Karkeasti, kaksi objektia ovat lähellä toisiaan jos voimme merkittävästi 'pakata' yhden objektin toisen objektin tiedoilla. Ideana se, että jos kaksi palaa ovat samankaltaisempia, niin voimme ytimekkäästi kuvailla yhden kun toinen on annettu. Tämän esittelemme luvussa 2 ja esittelemme mihin teoriaan algoritmi perustuu ja miten se toimii. Edellä mainitun vakauden esittämiseen käytämme useaa tosielämän pakkausalgoritmi': tilastollista (PPMZ), Lempel-Ziv -algoritmiin pohjautuvaa hakemistoa (gzip), lohkoperusteista (bzip2) tai erityistä tarkoitusta varten (Gencompress).

Luvussa 3 esittelemme algoritmin käyttökohteita monelta eri alueelta. Aloitamme sillä, miten yleiseti NCD:n avulla pystymme klusteroimaan tuloksia eri kategorioihin; miten musiikkikappaleet klusteroituu saman artistin alle, miten kuvantunnistuksessa saamme ryhmitettyä samankaltaiset kuvat ja miten sienten genomista saamme tarkan lajiryhmityksen.

Syvennymme musiikin, kuvantunnistuken ja dokumenttien kategorisoinnin tuloksiin lopussa lukua.

Luvussa 4 esitellään NCD:n kestävyyttä ja ongelmia. Ensiksi esitellään miten kohina vaikuttaa NCD:n tuloksiin lisäämällä sitä vähitellen toiseen niistä tiedostoista jota pakataan. Saamme nähdä miten paljon kohina vaikuttaa NCD:n laskemiin etäisyyksiin ja että vaikuttaako se klusterointiin ollenkaan.

Mikään algoritmi ei ole täydellinen ja niin on tämänkin kanssa. Algoritmissä itsessään ei ole selvää heikkoutta, mutta sen käytössä on otettava

kompressorin valinta huomioon kun halutaan suorittaa klusterointia. Monet suosituista pakkausalgoritmeistä ovat optimoituja tietyn kokoisille tiedostoille, niissä on niinkutsuttu ikkunakoko(engl. window size) joka määrittelee mikä tiedostokok on sopiva. Jos tiedostonkoko on pienempi kuin ikkunakoko, niin pakkaus on tehokasta, kun mennään sen yli niin pakkauksesta tulee huomattavasti tehottomampaa. Tässä luvussa esitellään tuloksia eri pakkausalgoritmejen vertailuista ja mikä näistä algoritmeistä on parhaimmaksi havaittu NCD:n kanssa käytettäväksi.

NCD ei ole ainut metriikka jolla voidaan mitata samankaltaisuutta. Internettiä hyödyntäen on tehty metriikka joka käyttää hakukoneita samankaltaisuuden tutkimiseen, tämä on nimetty Google samankaltaisuusetäisyydeksi (engl. Google Similarity Distance); se toimii myös muilla hakukoneilla kuin Googlella. Luvussa 5 esitellemme tämän sekä muita samankaltaisuuden metriikoita.

2 Normalisoitu Pakkausetäisyys: Mistä se koostuu, miten se toimii?

2.1 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.

2.2 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 [CVdW04]

$$E(x, y) = max\{K(x|y), K(y|x)\}. \quad (1)$$

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

$$NID(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}}. \quad (2)$$

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.

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

2.4 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

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}. \quad (3)$$

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

$$NCD(x,y) = \frac{C(xy) - C(x)}{C(y)}. \quad (4)$$

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.

3 Käyttökohteet

3.1 Klusterointi

3.1.1 Tuloksia

3.2 Kuvantunnnistus

4 Algoritmin ongelmat ja ominaisuudet

4.1 Kohinansietokyky

Kun NCD:tä käytetään kahteen eri tiedostoon toista näistä voi pitää kohinallisena versiona ensimmäisestä. Progressivisen kohinan lisääminen tiedostoon

voi tuottaa tietoa mittarista(measure) itsestään. Tämän vastaavuuden perusteella voimme tehdä teoreettisen päätelmän odotetusta kohinan lisäämisen vaikutuksesta algoritmiin, mikä selittää miksi NCD voi saada suurempia arvoja kuin 1 joissain tapauksissa. [CAO07]

4.2 Kompressorin valinta

5 Muita samankaltaisuuden metriikoita

5.1 Google Similarity Distance

Lähteet

[CAO07] Cebrian, M., Alfonseca, M. ja Ortega, A.: *The Normalized Compression Distance Is Resistant to Noise*. Information Theory, IEEE Transactions on, 53(5):1895–1900, 2007, ISSN 0018-9448.

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

[CVdW04] Cilibrasi, Rudi, Vitanyi, Paul ja Wolf, Ronald de: *Algorithmic Clustering of Music*. Web Delivering of Music, International Conference on, 0:110–117, 2004.