

Informatsioonikaugus

Mart Sõmermaa

4. november 2003. a.

Kokkuvõte

Käesolevas artiklis antakse ülevaade sõne Kolmogorovi keerukuse mõistel põhinevast universaalsest meetrikast, nn *informatsioonikaugusest*, ja selle praktilistest rakendustest geneetikas, lingvistikas ning autorsuse tuvastamisel.

1 Sissejuhatus

Objektide sarnasuse määramine on andmekaevanduse üks fundamentaalprobleeme, sarnasuse mõistel põhinevad nii otsingu- kui klassifitseerimisalgoritmide. Ming Li, Paul Vitányi *et al.* esitavad artiklites [LCL⁺03, BGL⁺98] üldise sarnasuse teooria, näidates, et sõne informatsioonisisalduse (Kolmogorovi keerukuse) abil saab defineerida universaalse meetrika, nn *informatsioonikauguse*. Universaalsus tähendab antud kontekstis seda, et kui kaks objekti on sarnased mingi suvalise meetrika põhjal, siis on need vähemalt sama sarnased informatsioonikauguse põhjal.

Ilmneb, et kirjeldatud meetrika on ka praktikas hästi kasutatav, kuigi, kuna sõne x Kolmogorovi keerukus $K(x)$ ei ole arvutatav, kasutatakse rakendustes tihendamist kui $K(x)$ heuristilist lähendit.

2 Kaugus

Kaugus (meetrika) on mingi hulga M otseruudul määratud mittenegatiivne funktsionaal $d : M \times M \rightarrow \mathbb{R}$, mis iga $x, y, z \in M$ korral rahuldab nn Fréchet' aksiome:

1. $d(x, y) = 0$ parajasti siis, kui $x = y$ (*samasusaksioom*),
2. $d(x, y) = d(y, x)$ (*sümmeetriaaksioom*),

3. $d(x, y) \leq d(x, z) + d(z, y)$ (kolmnurgaaksioom).

Hulka M , millel on defineeritud kaugus, nimetatakse meetriliseks ruumiks. Edaspidises näidatakse, et paar $(\{0, 1\}^*, d)$, kus d on informatsioonikaugus, on meetriline ruum.

Levinuimad üldised meetrikad:

1. Boole'i kaugus, on defineeritud igal hulgal

$$\begin{cases} d_B(x, y) = 0, & \text{kui } x = y, \\ d_B(x, y) = 1, & \text{vastasel juhul;} \end{cases}$$

2. kauguste pere $L_m : K^n \times K^n \rightarrow \mathbb{R}$, $m \in \mathbb{N}$, kus K^n on vektorruum üle korpuse K , erinevate normide suhtes,

$$L_m(x, y) = \|x - y\|_m = \left(\sum_{i=1}^n |x_i - y_i|^m \right)^{\frac{1}{m}}.$$

Peamiselt pakuvad huvi juhud $m = 1$ (Manhattani kaugus) ja $m = 2$ (eukleidiline kaugus) ning erijuht L_∞ ,

$$L_\infty(x, y) = \|x - y\|_\infty = \max(|x_1 - y_1|, \dots, |x_n - y_n|).$$

3. hulkade kaugus $d_\Delta : 2^X \times 2^X \rightarrow \mathbb{N}$, kus X on mingi universaalne hulk. Kui $A \subset X$, $B \subset X$, siis

$$d_\Delta(A, B) = |A \Delta B|.$$

4. Hammingu kaugus $d_H : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \mathbb{N}$,

$$d_H(x, y) = \sum_{i=1}^n x_i \oplus y_i = \sum_{i=1}^n |x_i - y_i| = d_1,$$

seega on Hammingu kaugus Manhattani kauguse erijuht. Seades hulgale vastavusse tema karakteristikliku funktsiooni, saame ka vastavuse hulkade kauguse ja Hammingu kauguse vahel.

5. teisenduskaugus (Levenšteini kaugus) d_E on Hammingu kauguse üldistus, kui viimane on bittide arv, mida tuleb muuta, et teisendada kahendsõne x sõneks y , siis d_E on vähim lisamiste, eemaldamiste ja asenduste arv sõne x teisendamisel sõneks y .

Lisaks nimetatutele on olemas hulgaliselt spetsiifilisi kaugusi. Eraldi tuleks märkida, et ka varasemates töödes on püütud tihendamist sõnede kauguse hindamisel kasutada, enne käesolevate autorite töid pole aga esitatud kõiki Frechet' aksioome rahuldavat teoreetiliselt hästipõhjendatud meetrikat.

Paljudes rakendustes on loetletud kaugused piisavad objektide sarnasuse määramiseks. On aga rakendusi, kus oleks vaja üldisemat meetrikat — näiteks kahe must-valge pildi sarnasuse määramine. Olgu antud mingi pildi esitus kahendsõnena. Kui muuta sõnes üksikuid bitte, jääb nii Hammingu kui eukleidiline kaugus originaali ja modifikatsiooni vahel väikeseks. Samas, pildi negatiivkujutis on mõlema meetrika korral originaalist maksimaalselt kaugel, inimvaatleja suudab aga piltide sarnasust tuvastada. Sama probleem ilmneb bitinihke korral — originaali ja modifikatsiooni Hammingu kaugus võib olla väga suur, kuigi pildid on sarnased. Analoogilisi kauguse spetsiifikast tulenevaid probleeme on ka teistes valdkondades, muuhulgas bioinformaatikas.

Informatsioonikaugus hõlmab kõiki loetletud meetrikaid omamata paljude spetsiifiliste meetrikate puudusi, muuhulgas määrab ka eelmises lõigus kirjeldatud originaali ja negatiivi kauguse vastavalt inimvaatleja intuitsioonile.

3 Kolmogorovi keerukus

Põhjaliku ülevaate Kolmogorovi keerukusest ja informatsiooniteooriast annab Ming Li ja Paul Vitányi raamat [LV97]. Käesolevaga anname lühidalt edaspidises vajalikud tähistused ja mõisted. *Sõneks* nimetame edaspidi kahendjärjendit (so lõplikku kahendjada). Leiduvad loomulikud kujutused suvaliste teiste lõplike objektide hulgast sõnede hulka. Sõnede hulka tähistatakse $\{0, 1\}^*$. Sõne x *Kolmogorovi keerukus* või *algoritmiline entroopia* $K(x)$ on lühima kahendprogrammi pikkus, mis väljastab sõne x mingil antud universaalarvutil (näiteks Turingi masinal). Kuigi funktsioon K on defineeritud antud masina jaoks, on see Churchi teesi põhjal mingi konstantse liidetavani masinasõltumatu ja universaalne. x^* tähistab lühimat programmi, mis väljastab sõne x , seega $|x^*| = K(x)$, kus $|s|$ tähistab sõne s pikkust.

Sõne x *tinglik Kolmogorovi keerukus* sõltuvalt sõnest y defineeritakse analoogiliselt kui lühima programmi pikkus, mis väljastab sõne x sisendsõne y korral, tinglikku keerukust tähistatakse $K(x|y)$. Sümboliga $K(x, y)$ tähistatakse lühima kahendprogrammi pikkust, mis väljastab sõned x ja y ning kirjelduse, kuidas neid eristada.

Definitsioon 1. *Funktsioon $f(x)$ on*

- ülalt arvatav¹, kui leidub rekursiivne funktsioon $g(x, t)$ selliselt, et

$$g(x, t + 1) \leq g(x, t) \quad \text{ja} \quad \lim_{t \rightarrow \infty} g(x, t) = f(x),$$

- alt arvatav, kui $-f$ on ülalt arvatav ning
- arvatav, kui ta on nii ülalt kui alt arvatav.

Lihtne on näha, et funktsioonid $K(x)$ ja $K(y | x^*)$ on ülalt arvatavad ja saab tõestada, et need ei ole arvatavad. Sõnes y sisalduv informatsioon sõne x kohta defineeritakse järgmiselt

$$I(x : y) = K(x) - K(x | y^*).$$

Saab näidata, et mingi konstantse liidetavani $I(x : y) \stackrel{\pm}{=} I(y : x)$, st

$$K(x) + K(y | x^*) \stackrel{\pm}{=} K(y) + K(x | y^*). \quad (1)$$

Sümboleid $\stackrel{\pm}{=}$ ja $\stackrel{\log}{=}$ kasutame edaspidi tähistamaks võrdust vastavalt mingi konstantse ja logaritmilise liidetavani.

4 Informatsioonikaugus

Sõnede x ja y vaheline *informatsioonikaugus* on lühima kahendprogrammi pikkus, mis väljastab sisendsõne x korral sõne y ja vastupidi. Kuna programm on lühim, kasutab see ära kogu andmeliiasuse mõlemas suunas teisendamisel. Nõutakse, et programm teisendamise käigus ei muutuks. Artiklis [BGL⁺98] näidatakse, et mingi logaritmilise liidetavani võrdub informatsioonikaugus

$$E(x, y) \stackrel{\log}{=} \max\{K(y | x), K(x | y)\}. \quad (2)$$

Kuna tinglik Kolmogorovi keerukus on ülalt arvatav, on ka informatsioonikaugus ülalt arvatav.

5 Normaliseeritud kaugus

Aktsepteeritavate meetrikate klassi defineerimisel tuleks välistada ebarealistlikud kaugusmõõdud nagu $\hat{d}(x, y) = \frac{1}{2}$, $\forall x \neq y$, piirates objektide hulka

¹ upper semi-computable

fikseeritud objekti mingis ümbruses. Seetõttu vaadeldakse edaspidi ülalt ar-
vutatavaid meetrikaid $D(x, y)$, mis rahuldavad tiheduse tingimust

$$\sum_{y \neq x} 2^{-D(x, y)} \leq 1. \quad (3)$$

Pikad sõned, mis erinevad n biti võrra, on intuiitiivselt sarnasemad kui lühikesed sõned, mis erinevad samuti n biti võrra. Seetõttu normaliseeritakse sõnekaugus järgnevalt:

olgu $D(x, y)$ ülalt arvutatav kaugus, mis rahuldab tingimust (3); olgu $n(D, x, y) = d_{n, D}(x, y)$ funktsioon, mille väärtused kuuluvad vahemikku $[0, 1]$ ja mis rahuldab normaliseerimistingimust

$$\sum_{y \neq x} 2^{-d_{n, D}(x, y)K(x)} \leq 1.$$

Olgu $K(x) = k$ ja $d \in [0, 1]$, siis

$$|\{y : d_{n, D}(x, y) \leq d, K(y) \leq k\}| \leq 2^{dk}. \quad (4)$$

Definitsioon 2. Normaliseeritud sõnekaugus või sarnasusmeetrika on meetrika $m(x, y)$, $x, y \in \{0, 1\}^*$, mille väärtused kuuluvad vahemikku $[0, 1]$ (protsessis $\max\{K(x), K(y)\} \rightarrow \infty$ hääbuva veaga) ja mis rahuldab tiheduse tingimust (4).

6 Normaliseeritud informatsioonikaugus

Artiklis [LBX⁺01] esitas sama töögrupp esialgse normaliseeritud informatsioonikauguse definitsiooni:

Definitsioon 3. Olgu x, y suvalised järjendid. Defineerime funktsiooni

$$d_s(x, y) = \frac{K(x | y^*) + K(y | x^*)}{K(x, y)} \quad (5)$$

Võrduse (1) põhjal saame

$$d_s(x, y) = 1 - \frac{K(x) - K(x | y^*)}{K(x, y)},$$

kus $K(x) - K(x | y^*)$ on sõnede ühine informatsioon $I(y : x)$. See kaugus rahuldab kolmnurga võrratust mingi liidetava veaga ja universaalsustingimust (defineeritakse edaspidi) mingi konstantse tegurini² $c < 2$.

Matemaatilisel täpsem ja adekvaatsem on järgnev definitsioon.

²saab näidata, et $d_s(x, y) \leq cD(x, y)$, $c < 2$, kus D on suvaline ülalt arvutatav kaugus.

Definitsioon 4. Olgu x, y suvalised järjendid. Defineerime funktsiooni

$$d(x, y) = \frac{\max\{K(x | y^*), K(y | x^*)\}}{\max\{K(x), K(y)\}} \quad (6)$$

$d(x, y)$ loomulik interpretatsioon: kui $K(y) \geq K(x)$, siis

$$d(x, y) = \frac{K(y) - I(x : y)}{K(y)}.$$

St kaugus $d(x, y)$ sõnede x ja y vahel on mittejagatud informatsioonikoguse suhe maksimaalsesse võimalikku jagatud informatsioonikogusesse.

On ilmne, et $d(x, y)$ on sümmeetriline ja rahuldab samasusaksioomi,

$$d(x, x) = O\left(\frac{1}{K(x)}\right).$$

Saab näidata, et $d(x, y)$ rahuldab mingi hääbuva veaga kolmnurgaaksioomi:

$$d(x, y) \leq d(x, z) + d(z, y) + O\left(\frac{1}{\max\{K(x), K(y), K(z)\}}\right),$$

ning normaliseerimistingimust (4). Seega kehtib järgnev teoreem:

Teoreem 1. Funktsioon $d(x, y)$ on normaliseeritud informatsioonikaugus.

7 Universaalsus

Ilmneb, et meetrika $d(x, y)$ hõlmab kõiki arvutatavaid sarnasusmõõte — kui kaks objekti on sarnased (normaliseeritud jagatud informatsioonikoguse mõttes) suvalise arvutatava meetrika järgi, siis on need objektid vähemalt sama sarnased meetrika $d(x, y)$ järgi:

Teoreem 2. Normaliseeritud informatsioonikaugus $d(x, y)$ on mingi protsessis $\max\{K(x), K(y)\} \rightarrow \infty$ hääbuva veaga asümptootiliselt ekvivalentne või asümptootiliselt kõrgemat järku mistahes ülalt arvutatava normaliseeritud meetrika $f(x, y)$ suhtes,

$$d(x, y) \leq f(x, y) + O\left(\frac{\log k}{k}\right), \text{ kus } k = \max\{K(x), K(y)\}.$$

Teoreemi tõestus on esitatud artiklis [LCL⁺03].

8 Rakendused

Kuigi meetrika d on teoreetiliselt hästi põhjendatud ja universaalne, ei ole $K(x)$ ja seega ka d arvutatav. Intuitiivselt on mõistetav, et *tihendamine* on $K(x)$ heuristiline lähend. Rakendustes kasutataksegi valdkonnaspetsiifilisi tihendusalgoritme meetrikate d_s ja d lähendite leidmisel. Meetrikad tuleb esmalt sobivale kujule viia:

seose (1) põhjal $K(x|y) \stackrel{\pm}{=} K(x, y) - K(y)$ ning on lihtne näidata, et $K(x, y) \stackrel{\log}{=} K(xy)$, kus xy tähistab sõnade x ja y konkatenatsiooni. Seega

$$d_s(x, y) = 1 - \frac{K(x) - K(x|y)}{K(x, y)} \approx 1 - \frac{K(x) + K(y) - K(xy)}{K(xy)},$$

ning kasutades tihendusalgoritmi $C(s)$ kui $K(s)$ heuristilist lähendit, $K(s) \approx |C(s)|$, saame

$$d_s(x, y) \approx 1 - \frac{|C(x)| + |C(y)| - |C(xy)|}{|C(xy)|}. \quad (7)$$

Analoogiliselt, eeldades üldisust kitsendamata, et $|C(x)| \leq |C(y)|$,

$$d(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \approx \frac{|C(xy)| - |C(x)|}{|C(y)|}. \quad (8)$$

Heuristiliste meetrikate (7) ja (8) abil leitakse uuritava sõnade hulga sarnasusmaatriks, mida kasutatakse otseselt järelduste tegemiseks (nt plagiaatide tuvastamine) või sisendina teistele algoritmidele (nt klasterdamine).

8.1 Näiteid rakendustest

Informatsioonikaugus on osutunud võimsaks tööriistaks, mida on kasutatud probleemide lahendamisel väga erinevates valdkondades:

1. genoomide võrdlemine bioinformaatikas. Genoomide sarnasusmaatriksi leidmine informatsioonikauguse abil on täisautomaatne ning seejuures ei ole vaja gene eraldi tuvastada.

Artiklites [LCL⁺03, LBX⁺01] antakse ülevaade imetajate filogeneesipuu tuletamisest informatsioonikauguse abil. Genoomi Kolmogorovi keerukuse hindamiseks kasutati algoritmi GENCompress, mis artikli [CKL00] põhjal annab parima tulemuse geeni-informatsiooni tihendamisel; puu konstrueerimisel kasutati naabriliite (*neighbour joining*) meetodit paketi MOLPHY [AH96].

Imnes, et primaadid, sh inimene, on lähemal imetajate hõimkonnale *Ferungulata* (sõralised, kabjalised, londilised ja kiskjad) kui närilistele.³

Paul Vitányi kirjeldab, kuidas sama meetodiga klassifitseeriti koheselt viirus SARS (<http://homepages.cwi.nl/~paulv/papers/sarsvirii>).

2. autorsuse tuvastamine. Artiklis [CLMS02] kirjeldatakse lähteteksti-plagiaatide tuvastamise süsteemi SID, mis põhineb informatsioonikaugusel, ning demonstreeritakse, et levinud plagiaadituvastusprogrammide petmise tehnikad süsteemi SID tulemusi ei mõjuta.

Artiklites [CVdW03, Mui03] tutvustatakse muusikateoste autorsuse ja žanri automaatse tuvastamise süsteemi, kus informatsioonikauguse lähendina kasutatakse tihendusalgoritmi *bzip2*.

3. keelte suguluse määramine lingvistikas. Artiklis [BGL⁺98] antakse ülevaade informatsioonikauguse kasutamisest keeltepuu automaatsel konstrueerimisel inimõiguste ülddeklaratsiooni tõlgetest 52 eri keelde. Ootuspäraselt paigutus inglise keel romaani keelte hulka, kuna teatavasti on selles hulgaliselt prantsuse laensõnu, ning ungari keel ei paigutunud soomeugri keelte hulka, kuna seda on ulatuslikult mõjutanud slaavi ja türki keeled. Ülejäänud keeled paigutusid vastavalt üldlevinud lingvistilistele teooriatele. Informatsioonikauguse määramisel kasutati tihendusalgoritmi *gzip*, klassifikatsioonipuu koostamisel Fitch-Margoliashi meetodit [FM67] paketi PHYLIP [AH96].

Vastukaja (näiteks artikkel [Goo02]) tekitas Itaalia lingvistide töögrupi artikkel [BCL02], kus samuti kasutati tihendamisel põhinevaid meetodeid keelepuu tuletamiseks ning autorsuse tuvastamiseks. Kuigi töös viidati informatsioonikaugusele, kasutati *ad hoc* meetodeid, mis ei olnud teoreetiliselt põhjendatud ning töös defineeritud “kaugus” ei rahuldanud Frechet’ aksioome.

9 Kokkuvõte

Käesolevas artiklis näidati, et informatsioonikaugus on teoreetiliselt hästipõhjustatud universaalne meetrika, mille abil on võimalik lahendada probleeme väga erinevates valdkondades. Kuigi tihendamisel põhinevaid meetodeid on sõnade sarnasuse määramisel kasutatud ka varem, pole enne esitatud kõiki Frechet’ aksioome rahuldavat üldist meetrikat.

³bioloogias on üheselt lahendamata probleem, kas geeni-info põhjal tuleks imetajaid grupeerida (*(Primates, Rodentia), Ferungulata*) või (*(Primates, Ferungulata), Rodentia*).

Informatsioonikaugus defineeriti lõplikult alles 2003. a. Seega on tegemist uue meetrikaga, mille rakendusvaldkonnad kindlasti ei piirdu artiklis loetletutega. Informatsioonikaugusel põhinevaid süsteeme võib suvalises valdkonnas kasutada andmekaevandusautomaatidena, mis tuvastavad iseseisvalt (ilma inimsekkumiseta) seniavastamata sarnasusi andmelattu talletatud objektide vahel; seda on edukalt tehtud plagiaatide tuvastamisel, lingvistikas ja bioinformaatikas.

Viited

- [AH96] Jun Adachi and Masami Hasegawa. Molphy: A computer program package for molecular phylogenetics, 1996.
- [BCL02] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. Language trees and zipping. *Physical Review Letters*, 88(048702), 2002.
- [BGL⁺98] Charles H. Bennett, Peter Gacs, Ming Li, Paul M. B. Vitányi, and Wojciech H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.
- [CKL00] Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences and its applications in genome comparison. In *Proceedings of the fourth annual international conference on Computational molecular biology*, page 107. ACM Press, 2000.
- [CLMS02] Xin Chen, Ming Li, Brian Mckinnon, and Amit Seker. A theory of uncheatable program plagiarism detection and its practical implementation, May 2002. Käsikiri.
- [CVdW03] Rudi Cilibrasi, Paul M. B. Vitányi, and Ronald de Wolf. Algorithmic clustering of music, 2003.
- [FM67] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–84, Jan 1967.
- [Goo02] Joshua Goodman. Extended comment on language trees and zipping, 2002.
- [LBX⁺01] Ming Li, Jonathan H. Badger, Chen Xin, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

- [LCL⁺03] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 863–872. Society for Industrial and Applied Mathematics, 2003.
- [LV97] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, second edition, 1997.
- [Mui03] Hazel Muir. Software to unzip identity of unknown composers. *New Scientist*, April 2003.