

MDL — lühima kirjelduse printsiip

Meelis Kull
Meelis.Kull@ut.ee

Andmekaevandamise uurimisseminar MTAT.03.169.
Arvutiteaduse instituut, Tartu Ülikool
Detsember 2003, lk. 51–69

Kokkuvõte

Masinõppes kasutatakse palju induktiivset järeldamist (ik *inductive inference*), kus järeldus (ehk mudel) tehakse ebapiisava andmehulga alusel ning kehtib seega ainult mingi teadmata „tõenäosusega”. Lühima kirjelduse printsiip (MDL) on üks võimalus hinnata, milline mudel mingist vaadeldavate mudelite hulgast on parim lähtuvalt olemasolevatest andmetest. Käesoleva artikli eesmärk on anda ülevaade MDL-i kahest versioonist — kaheosalisest MDL-ist ja stohhastilisel keerukusel baseeruvast MDL-ist. Uurimusliku osana vaadeldakse bitistringide perioodilisi mudeleid ning rakendatakse MDL-printsiipi parima perioodilise mudeli valimiseks.

1. Sissejuhatus

Andmetest teadmiste kogumine hõlmab kaht olulist protsessi — deduktiivset ja induktiivset järeldamist. Deduktiivne järeldamine on loogikareeglite rakendamine ning seeläbi ei teki midagi semantiliselt uut, mida ei oleks algandmetes¹. Induktiivne järeldus tehakse puuduliku informatsiooni alusel ning tulemus on tõene vaid tõenäoliselt, kusjuures see tõenäosus on meile teadmata. Näiteks kui mingi funktsiooni $f : \mathbb{N} \rightarrow \mathbb{N}$ korral $f(x) = x$ iga tuhandest väiksema x korral, siis

¹Deduktiivse ja induktiivse järeldamise definitsioonid on toodud järgmistel veebilehtedel:
http://www.psych.ualberta.ca/~mike/Pearl_Street/Dictionary/contents/D/deductive_inference.html
http://www.psych.ualberta.ca/~mike/Pearl_Street/Dictionary/contents/I/inductive_inference.html.

„tõenäoliselt” $f(1000) = 1000$. Samas ei saa me välistada, et $f(1000) = 1001$, kui funktsioon f juhtub olema defineeritud näiteks võrdusega $f(x) = x + \lfloor \frac{x}{1000} \rfloor$. Kumb hüpotees on parem — kas $f(1000) = 1000$ või $f(1000) = 1001$? Occam’i habemenoa printsiibi (ik *Occam’s razor*) järgi tuleks juhul, kui hüpoteeside tree-ningandmetega kokkusobivus on sama, eelistada kõige lihtsamat hüpoteesi (vt. näiteks (Domingos 1998)). Antud juhul on ilmselt lihtsaim hüpotees $f(x) = x$, ehk siis $f(1000) = 1000$. Occam’i habemenoa printsiibi rakendamisele sellisel kujul on olnud ka vastuväiteid (Domingos 1998), kuid sellest hoolimata rakendatakse seda praktikas laialdaselt. Occam’i habemenoa printsiipi kasutab ka lühima kirjelduse printsiip (ik *Minimum Description Length principle*, MDL).

Andmekaevanduses tuleb samuti teha induktiivseid järeldusi — näiteks kui ülesandeks on koostada mudel, mis kirjeldab võimalikult hästi olemasolevaid andmeid, kusjuures meil ei ole teada andmete tekkimise mehhanism (õige mudel). Edaspidises eeldame, et meil on juba olemas mingi hulk mudeleid, mille seast tuleb kõige parem välja valida. Mudeli valimisel võivad kergesti tekkida kas alasobitamine (ik *underfitting*) või ülesobitamine (ik *overfitting*). Alasobitamise puhul ei kirjelda mudel andmeid piisavalt hästi, ülesobitamine tähendab aga seda, et mudel kirjeldab küll väga hästi antud andmeid, kuid on väga keeruline (näiteks keerulisem kui andmed ise). Enamasti on otsitav mudel kompromiss lihtsuse ja kirjeldustäpsuse vahel. Tihti juhtub, et andmete kohta teadaoleva valdkonna-spetsiifilise informatsiooni alusel ei saa lõplikult otsustada, milline mudel on parim. Üks võimalus on sellisel juhul kasutada informatsiooniteoreetilist lähenemist, mille järgi tuleks valida mudel, mis annab andmetele lühima kirjelduse (Hansen & Yu 2001). Sellel peatumegi edaspidises, tuues selleks kõigepealt sisse andmekogumi Kolmogorovi keerukuse mõiste. Nimelt on andmete pakkimine Kolmogorovi keerukuse mõttes peaaegu alati parim meetod (Grünwald 1998). Kuna aga Kolmogorovi keerukuse leidmine osutub mittelahenduvaks ülesandeks, siis siirdume ühe praktikas rakendatava meetodi, lühima kirjelduse printsiibi (edaspidi MDL) juurde. Esmalt tutvume ühe näite varal MDL-i lihtsama, kaheosalise versiooniga, mille järgi on parim mudel see, mis minimiseerib mudeli kirjeldamiseks ja mudeli alusel andmete kirjeldamiseks kuluva bittide arvu summa. Seejärel defineerime tõenäosusliku mudeli mõiste ning kirjeldame selle seost koodidega. Tõenäosuslikud mudelid võimaldavad meil leida mudelile vastav kood andmete kodeerimiseks. Viimasena vaatleme üheosalist versiooni MDL-printsiibist, defineerides enne selleks andmete stohhastilise keerukuse mudeliklassi suhtes.

parima Turingi masina leidmisest. Küll aga on sarnast ideed võimalik praktikas rakendada, selle juurde järgnevalt siirdumegi.

3. MDL-printsiibi kaheosaline versioon

Kolmogorovi keerukus annaks andmetele küll häid mudeleid, kuid selle leidmise mittelahenduvuse tõttu pole Kolmogorovi keerukust võimalik praktikas rakendada. Jorma Rissanen pakkus aastal 1978 välja lühima kirjelduse printsiibi (ik *Minimum Description Length principle*, MDL), mis on sarnane Kolmogorovi keerukusega, kuid on praktikas hästi rakendatav (Rissanen 1978). MDL-ist on mitmeid versioone, kõige lihtsam ja intuiitsem nendest on kaheosaline versioon. MDL-i kaheosalise versiooni järgi on kõige parem mudel see, mis minimiseerib mudeli kirjeldamiseks ja mudeli alusel andmete kirjeldamiseks kuluva bittide arvu summat. Kirjeldamine tähendab sisuliselt kodeerimist ning seda mõistet kohe käsitlemegi.

3.1. Kodeerimine

Arvu 1000000 võib kirjeldada ehk kodeerida mitmeti — näiteks matemaatilises keeles stringina „10⁶” või eesti keeles stringina „Miljon.”. Viimane on hea selle poolest, et kirjelduse lõpp on selgelt märgistatud punktiga. Selliseid kirjeldusi on võimalik ilma segadust tekitamata üksteise järele kirjutada, näiteks „Miljon.Miljon.”. Samal ajal ei saa kirjutada „10000001000000” või „10⁶10⁶” ilma, et tähendus muutuks. Kirjelduste lõpetatust nõuame kogu käesoleva töö jooksul, selle garanteerib järgnev definitsioon.

Definitsioon 1 Üksühest kujutust $C : A \rightarrow B^*$ nimetatakse prefikskoodiks hulgal A koodsõnade hulgaga B^* , kui ükski koodsõna ei ole ühegi teise koodsõna prefiksiks ehk alguseks.

See tähendab, et suvalise $a \in A$ korral pole koodsõnast $C(a)$ võimalik lõpust hulga B elementide ära jätmisel saada mingit muud koodsõna $C(a')$, kus $a' \in A$. Koodsõna on lõpetatud kirjeldus, sest ükskõik, mis sellele ka ei järgneks, me ei saa mingit muud koodsõna. Märgime veel, et käesolevas töös vaatleme ainult juhitud, kus koodsõnad on bitistringid ehk siis $B = \{0, 1\}$. Prefikskoodi nimetame edasises ka lihtsalt koodiks.

Näide 1 Olgu $A = \{a_1, a_2, a_3\}$. Kui defineerida

$$\begin{aligned} C(a_1) = 0, & \quad C(a_2) = 10, & \quad C(a_3) = 11 & \quad \text{ja} & \quad C(a_4) = 101; \\ C'(a_1) = 0, & \quad C'(a_2) = 10, & \quad C'(a_3) = 110 & \quad \text{ja} & \quad C'(a_4) = 111, \end{aligned}$$

siis C pole kood, sest $C(a_4)$ on $C(a_2)$ prefiksiks. Küll aga sobib prefiks-koodiks kujutus C' .

Viimaks toome veel sisse koodipikkusfunktsiooni mõiste, mis hakkab hiljem olulist rolli mängima.

Definitsioon 2 Elemendi $a \in A$ kodeerimiseks kuluvate hulga B elementide arvu tähistame $L(a)$. Vajadusel täpsustame ja märgime ära kodeerimisel kasutatava koodi C , kirjutades $L_C(a)$. Funktsiooni $L : A \rightarrow \mathbb{N}$ nimetame koodipikkus-funktsiooniks.

3.2. Kaheosaline MDL

Tähistagu D andmeid, mida vaatleme lõpliku järjendina mingi hulga E elementidest, $D \in E^*$. Hulka E^* nimetame edaspidi *andmeruumiks*. Näiteks võib (kuid ei pruugi) olla $E = \{0, 1\}$. Järgnevas eeldame lihtsuse mõttes, et E on lõplik. Põhimõtteliselt jääb kogu edaspidine teooria kehtima ka juhul, kui E on loenduv või $E \subset \mathbb{R}^k$. Ainult et esimesel juhul muutuvad lõplikud summad ridade summadeks ning teisel juhul integraalideks. Samuti eeldame, et andmete pikkus n on fikseeritud ja andmeruum on seega E^n . Ka see eeldus pole hädavajalik, kuid lihtsustab käsitlust. Üldise juhu kohta vt. näiteks (Grünwald 1998).

Vaadeldavate mudelite hulka tähistame \mathcal{M} — nende hulgast tuleb meil välja valida parim mudel. Nagu enne juba mainitud, on MDL-i kaheosalise versiooni järgi kõige parem mudel see, mis minimiseerib mudeli kirjeldamiseks ja mudeli alusel andmete kirjeldamiseks kuluva bittide arvu summat.

MDL ei määra ära, milliseid koode kirjeldamisel kasutatakse. Selles mõttes ei ole MDL universaalne. Küll aga on olemas meetodid, kuidas saada suhteliselt loomulikke koode, ühte käsitleme ka siin edaspidi. Tähistame nüüd ära MDL-i jaoks olulised koodid:

- $C : \mathcal{M} \rightarrow B^*$ — kood, millega kodeeritakse mudeleid;
- $C_M : E^n \rightarrow B^*$, iga $M \in \mathcal{M}$ jaoks — kood, millega kodeeritakse mudeli M alusel andmeid.

MDL-i mõttes parim on mudel $M_{2\text{-MDL}}$, mille korral MDL-hinnang $L_C(M) + L_{C_M}(D)$ on minimaalne, ehk teisisõnu:

$$M_{2\text{-MDL}} = \operatorname{argmin}_{M \in \mathcal{M}} (L_C(M) + L_{C_M}(D)).$$

Nüüd vaatleme üht näidet MDL-printsibi kaheosalise versiooni rakendamise kohta. Siinkohal algav näide kordub läbi käesoleva artikli ning on autori poolt tehtud uurimuslik osa.

Näide 2 Olgu ülesandeks ilma modelleerimine ühe kuu jooksul, arvutuste lihtsustamiseks võtame päevade arvuks $n = 32$. Oletame, et meil on tehtud ühe kuu jooksul ilmavaatlusi ning iga päeva kohta on fikseeritud, kas oli selge või vihmane (vastavalt 0 või 1). Olgu vaatluste tulemuseks bitistring

$$D = 01010100010100010100010101010100.$$

Andmemudelite hulgana \mathcal{M} vaatleme kõigi kuni 32-bitiste stringide hulka, kusjuures mudel $M \in \mathcal{M}$ väidab bitistringi M perioodilist kordumist andmetes D . Näiteks mudel $M = 101$ väidab, et andmeteks on

$$101|101|101|101|101|101|101|101|101|101|10.$$

MDL nõuab mudeli kirjeldamist (mis on antud juhul triviaalne — tuleb vaid esitada bitistring M) ning mudeli abil andmete kirjeldamist. Viimase puhul tuleb üles lugeda erandid — kohad, kus mudel ei kehti. Vaadeldavate andmete puhul paistab suhteliselt hästi sobivaks mudeliks olema $M = 01$. Mudeli ning andmete kirjeldused oleks siis järgmised.

mudel:	andmetes kordub alamstring 01
andmed:	eranditeks on 8., 14., 20. ja 32. bitt

$$D = 01|01|01|00|01|01|00|01|01|00|01|01|01|01|01|01|01|00$$

Selleks, et leida MDL-i mõttes parim mudel, peame kõigepealt fikseerima kasutatavad koodid. Mudel on juba bitistringina esitatud ning seega võiks koodina C kasutada samasusteisendust $C(M) = M$. Paraku pole tegemist koodiga, sest üks mudel võib olla teise prefiksiks. Et saada koodi, võime mudeli ette kirjutada tema pikkuse. Kuna pikkus võib olla 1 kuni 32, siis saab seda kodeerida viie bitiga (sest $2^5 = 32$). On lihtne veenduda, et nüüd oleme saanud koodi. Näiteks mudeli $M = 01$ korral kulub kodeerimiseks $L_C(M) = 5 + 2 = 7$ bitti.

Iga erand on arv vahemikust 1 kuni 32 ning seega saab seda kodeerida viie bitiga. Kood C_M võiks kodeerida erandid mudeli M kehtivuses näiteks järgmiselt.

0 <erand> 0 <erand> 0 ... 0 <erand> 0 <erand> 1

Eraldavad bitid on vajalikud selleks, et oleks aru saada, millal kirjeldus lõpeb. Kokku kulub siis bitte iga erandi kohta 6 pluss veel üks bitt kõige lõpus. Näiteks mudeli $M = 01$ korral kulub andmete D kodeerimiseks $L_{C_M}(D) = 4 \cdot 6 + 1 = 25$ bitti.

Kokkuvõttes kulub siis mudeli 01 korral mudeli ja andmete kodeerimiseks $7 + 25 = 32$ bitti. Nagu näha, kulus põhiline hulk bitte mitte mudeli, vaid erandite kodeerimiseks, seetõttu võiks proovida leida mudelit, millel on vähem erandeid. Järgmisel mudelil pole üldse erandeid.

mudel:	andmetes kordub alamstring 010101000101000101000101
erandid:	puuduvad

$D = 010101000101000101000101|01010100$

Selle mudeli korral kulub mudeli ja andmete kodeerimiseks kokku $L_C(M) + L_{C_M}(D) = (5 + 24) + (0 \cdot 6 + 1) = 30$ bitti, järelikult on MDL-printsiibi järgi see mudel parem kui eelmine. MDL-i järgi parim mudel peaks olema selline, milles on saavutatud väike erandite arv suhteliselt lühikese mudeliga. Antud näite puhul osutub parimaks järgmine mudel.

mudel:	andmetes kordub alamstring 000101
andmed:	eranditeks on 2. ja 26. bitt

$D = 0\underline{1}0101|000101|000101|000101|000101|0\underline{1}0101|00$

Siin kulub mudeli ja erandite peale kokku vaid $L_C(M) + L_{C_M}(D) = (5 + 6) + (2 \cdot 6 + 1) = 24$ bitti. Sama tulemuse annab ka mudel $M = 010101000101$, mille pikkus on 12 ning millel on üks erand.

Eelneva põhjal võib esitada MDL-i põhieesmärgi järgmiselt.

MDL-i eesmärk on leida mõistlik tasakaal reegli ja erandite vahel.

Nüüd me oleme ära kirjeldanud MDL-i kaheosalise versiooni põhiidee ning jäänud on vaid rääkida sellest, kuidas valida koode C ja C_M nii, et tulemus oleks võimalikult „hea”. Koodi C valimisel me käesolevas artiklis ei peatu, selle kohta võib lugeda allikatest (Grünwald 1998) ja (Hansen & Yu 2001). Järgnevas punktis hakkame kirjeldama ühte võimalust koodide C_M valimiseks. Kindlasti ei ole

see ainus võimalus, kuid piisavalt loomulik siiski. Me alustame mudelist M ning jõuame lõppkokkuvõttes välja koodini C_M . Ette rutates võib öelda, et tegelikult me väldime koodi C_M reaalsel konstrueerimist, meile piisab vaid selle koodipikkusfunktsiooni L_{C_M} teadmisest. Eesmärk on meil ju sellise mudeli $M \in \mathcal{M}$ valimine, mille korral $L_C(M) + L_{C_M}(D)$ on minimaalne.

Kõigepealt leiame väga loomulikult viisil tõenäosuslikule mudelile vastava koodi pikkuse ning seejärel vaatleme üldist, suvalise mudeli juhtu.

4. Tõenäosuslikud mudelid

Paljud huvitavad mudelite hulgad \mathcal{M} on tõenäosuslikud, s.t. et iga mudel $M \in \mathcal{M}$ esitab tõenäosusjaotust üle andmeruumi E^n . Teisisõnu, mudel M annab kõigile võimalikele andmetele $D = (x_1, \dots, x_n)$ mingi tõenäosuse, millega tekivad selle mudeli järgi just need andmed. Mida suurema tõenäosuse andmed mudeliga saavad, seda paremini sobivad need andmed mudeliga kokku.

Definitsioon 3 Funktsiooni $P : E^n \rightarrow [0, 1]$ nimetatakse tõenäosuslikuks mudeliks üle E^n , kui

$$\sum_{(x_1, \dots, x_n) \in E^n} P(x_1, \dots, x_n) = 1.$$

Näide 3 Vaatleme ilma modelleerimist ühe nädala jooksul ($n = 7$). Oletame, et vaatlustel oleme teinud sellised ligikaudsed tähelepanekud:

- esmaspäeval võib ilm olla selge või vihmane võrdselt tõenäosusega $\frac{1}{2}$;
- kui ilm on selge, siis järgmisel päeval on ilm selge või vihmane vastavalt tõenäosustega p ja $1 - p$;
- kui ilm on vihmane, siis järgmisel päeval on ilm selge või vihmane vastavalt tõenäosustega $1 - q$ ja q .

Selliste tähelepanekute järgi saame ilma kohta tõenäosusliku mudeli $P_{p,q}$. Kui näiteks $p = 0.8$ ja $q = 0.6$, siis

$$\begin{aligned} P_{p,q}(0000000) &= 0.5 \cdot 0.8^6 \approx 0.13 \text{ ja} \\ P_{p,q}(1010101) &= 0.5 \cdot (0.4 \cdot 0.2)^3 \approx 0.00026. \end{aligned}$$

Esimene neist on tõenäosus, et terve nädal on ilus ilm, ning teine tõenäosus, et vaheldumisi sajab ja on selge.

4.1. Tõenäosuslikud mudelid ja täpsed koodipikkusfunktsioonid

Intuitiivselt on selge, et kodeerimisel ei saa kõik koodsõnad tulla kuitahes lühikesed, on olemas mingi alumine piir. Näiteks ei saa olla koodi $C : \{a_1, a_2, a_3\} \rightarrow B^*$, nii et

$$L_C(a_1) = 1, \quad L_C(a_2) = 1 \quad \text{ja} \quad L_C(a_3) = 2.$$

Põhjus on selles, et siis oleks paratamatult kas $C(a_1)$ või $C(a_2)$ prefiksiks koodsõnale $C(a_3)$. Täpse alumise piiri seab Krafti võrratus, tõestust vt. (Grünwald 1998). Siin ja edaspidi kasutame tähistust $\mathbf{x} = (x_1, \dots, x_n)$.

Teoreem 1 (Krafti võrratus) *Kui C on kood hulgal E^n , siis*

$$\sum_{\mathbf{x} \in E^n} 2^{-L_C(\mathbf{x})} \leq 1.$$

Krafti võrratus ütleb sisuliselt, et lühikese koodipikkuse saab omistada ainult vähestele hulga E^n elementidele.

Lühima kirjelduspikkuse printsiibi järgi peaksime me kodeerima andmed võimalikult lühikeselt. Seega huvitavad meid sellised koodid, mis saavutavad Krafti võrratuse poolt määratud piiri.

Definitsioon 4 *Koodipikkusfunktsiooni $L : E^n \rightarrow \mathbb{N}$ nimetame täpseks, kui Krafti võrratuses kehtib võrdus. Koodi nimetame täpseks, kui tema koodipikkusfunktsioon on täpne.*

Lihtne on kontrollida, et näites 2 kasutatud kood C on täpne. Koodid C_M ei ole täpsed ning neid võiks proovida lühemaks teha.

Suvalisele täpsele koodile C vastab loomulikul viisil tõenäosuslik mudel, võrdusega $P(\mathbf{x}) = 2^{-L_C(\mathbf{x})}$. Tulemus on tõepoolest tõenäosuslik mudel, sest koodi täpsuse tõttu Krafti võrratuses kehtivast võrdusest saame, et kõikide E^n elementide tõenäosuste summa on 1.

Nüüd proovime suvalisest tõenäosuslikust mudelist P saada täpset koodi. Osutub, et kui defineerida koodipikkusfunktsioon L võrdusega $L(\mathbf{x}) = -\log P(\mathbf{x})$ ning kui $L(\mathbf{x})$ on täisarv iga $\mathbf{x} \in E^n$ korral, siis saame tõepoolest täpse koodi. Siin ja edaspidi tähistab \log kahendlogaritmi.

Teoreem 2 *Kui mingi funktsiooni $L : E^n \rightarrow \mathbb{N}$ korral kehtib Krafti võrratuses võrdus, siis leidub täpne kood sellise koodipikkusfunktsiooniga.*

Teoreemis mainitud täpset koodi saab leida näiteks Huffmani pakkimise algoritmiga (vt. näiteks (Kiho 2003)). Kui $L(\mathbf{x})$ ei ole aga täisarv, siis on võimalik tõestada, et leidub kood pikkusega $\lceil -\log P(\mathbf{x}) \rceil$, seda koodi nimetatakse Shannon-Fano koodiks (tõestust vt (Grünwald 1998)). Peaaegu sama lühikese koodipikkusfunktsiooniga kood saadakse ka Huffmani pakkimisel. Kuigi selle mittetäisarvulisuse probleemi tõttu ei pruugi leida koodi pikkusega $L(\mathbf{x}) = -\log P(\mathbf{x})$, vaatleme me sellist funktsiooni $L(\mathbf{x})$ ikkagi koodipikkusfunktsioonina, sest meie eesmärk on ju tegelikult leida lühima koodi pikkus, mitte kood ise.

Kokkuvõttes saime üksühese vastavuse tõenäosuslike mudelite ja täpsete koodipikkusfunktsioonide vahel. Selle vastavuse realiseerivad järgmised, omavahel samaväärsed seosed:

$$L(\mathbf{x}) = -\log P(\mathbf{x}) \iff P(\mathbf{x}) = 2^{-L(\mathbf{x})}. \quad (1)$$

Näide 4 Vaatleme taas ilma modelleerimist tõenäosuslike mudelitega $P_{p,q}$ (vt. näide 3). Lisaks teeme kitsendava eelduse $p, q \in \{0, \frac{1}{15}, \frac{2}{15}, \dots, \frac{14}{15}, 1\}$. Andmed olgu meil samad, mis näites 2,

$$D = 01010100010100010100010101010100.$$

Üritame nüüd MDL-printsiiibi kaheosalise versiooniga leida vaadeldavate mudelite seas sellist, mis vastaks kõige paremini andmetele D . Selleks peame kõigepealt fikseerima mudelite ja andmete kodeerimiseks kasutatavad koodid.

Mudeleid saame kodeerida 8 bitiga — nii p kui q kodeerimiseks kulub 4 bitti, sest neil on 16 võimalikku väärtust. Teisisõnu, $L_C(P_{p,q}) = 8$ kõigi mudelite $P_{p,q}$ korral.

Teiseks on meil vaja mudeli $P_{p,q}$ abil kodeerida andmed D . Siinkohal kasutame ära selle, et tõenäosuslikule mudelile $P_{p,q}$ vastab täpne koodipikkusfunktsioon $L_{C_{p,q}}$ seosega (1). Koodi ennast pole meil vaja, sest oluline on vaid koodi pikkus.

MDL-printsiiibi järgi on nüüd vaja leida mudel $P_{p,q}$, mis annab väikseima summa $L_C(P_{p,q}) + L_{C_{p,q}}(D)$. Et esimene liidetav on konstantselt võrdne arvuga 8, siis keskendume teise liidetava minimeerimisele. Valemist (1) saame, et $L_{C_{p,q}}(D) = -\log P_{p,q}(D)$, järelikult on meil vaja leida $P_{p,q}(D)$. Selleks paneme tähele, et andmetes D järgneb selgele ilmale selge ja vihmane ilm vastavalt 7 ja 12 korral ning vihmale ilmale järgneb kõigil 12 korral selge ilm. Kuna esimene päev on selge tõenäosusega $\frac{1}{2}$, siis kokku saame, et $P_{p,q}(D) = \frac{1}{2} \cdot p^7 \cdot (1-p)^{12} \cdot (1-q)^{12}$. Järelikult

$$\begin{aligned} L_{C_{p,q}}(D) &= -\log \frac{1}{2} - \log p^7 - \log(1-p)^{12} - \log(1-q)^{12} = \\ &= 1 - 7 \cdot \log p - 12 \cdot \log(1-p) - 12 \cdot \log(1-q). \end{aligned}$$

On lihtne kontrollida, et lubatud p ja q väärtuste seas annavad minimaalse tulemu-
se $p = \frac{6}{15}$, $q = 0$. Mudeli ja andmete kodeerimiseks kulub siis bittide kokku

$$8 + \left(1 - 7 \cdot \log \frac{6}{15} - 12 \cdot \log \left(1 - \frac{6}{15} \right) - 12 \cdot \log(1 - 0) \right) \approx 28.097.$$

5. Mudelist M koodini C_M

Eelmises punktis näitasime tõenäosusliku mudeli P jaoks ära koodi C_P , mida kasutada andmete kodeerimiseks kasutades seda mudelit. Käesoleva punkti eesmärk on leida kood C_M suvalise (mittetõenäosusliku) mudeli M jaoks. Selleks muudame mittetõenäosusliku mudeli tõenäosuslikuks ning seejärel leiame sellele vastava koodi. Meetod on kohandatud allikast (Grünwald 1998).

Mudeli tõenäosuslikuks muutmine on võimalik mitmeti ning seejuures peame tegema lisaeeldusi. Kõigepealt eeldame, et meil on lisaks mudelite hulgale \mathcal{M} ja andmeruumile E^n määratud reaalarvuliste väärtustega veafunktsioon $ER(M, D)$, mis näitab iga mudeli $M \in \mathcal{M}$ ja iga andmekogumi $D \in E^n$ korral, kui palju mudel M eksib andmetel D . Selge ja vihmade ilma perioodiliste mudelite näites 2 võiks veafunktsiooni väärtuseks $ER(M, D)$ olla erandite arv ehk siis nende päevade arv, kus mudel M ei klapi andmetega D . On mõistlik eeldada, et veafunktsiooni annab ette kasutaja, sest see annab mudelile tähenduse — siiani on mudel olnud lihtsalt üks element hulgas \mathcal{M} .

Teine eeldus puudutab vigade tekkimise tõenäosusi andmetes. Nimelt eeldame, et leiduvad konstandid $\alpha_1, \alpha_2 \in (0, 1]$, mille korral

$$\Pr[D|M] = \alpha_1 \cdot \alpha_2^{ER(M,D)} \quad \forall D \in E^n. \quad (2)$$

Hiljem veendume, et näites 2 kirjeldatud mudeliklassi \mathcal{M} korral on see eeldus tõepärane.

Võrduse (2) abil olemegi saanud algsest mittetõenäosuslikust mudelist M tõenäosusliku mudeli P , kusjuures $P(D) = \alpha_1 \cdot \alpha_2^{ER(M,D)}$. Kasutades eelmises punktis kirjeldatud seost (1) tõenäosuslike mudelite ja koodide vahel saame, et

$$L_{C_P} = -\log P(D) = (-\log \alpha_1) + (-\log \alpha_2) \cdot ER(M, D).$$

Selle saadud koodi valimegi mudelile M vastavaks koodiks C_M . Seega

$$L_{C_M}(D) = \beta \cdot ER(M, D) + K, \quad (3)$$

kus $K = -\log \alpha_1$ ja $\beta = -\log \alpha_2$. Et saadud kood on täpne, siis

$$1 = \sum_{D \in E^n} 2^{-L_{C_M}(D)} = \sum_{D \in E^n} 2^{-\beta \cdot \text{ER}(M,D) - K} = 2^{-K} \sum_{D \in E^n} 2^{-\beta \cdot \text{ER}(M,D)},$$

millest omakorda

$$K = \log \sum_{D \in E^n} 2^{-\beta \cdot \text{ER}(M,D)}. \quad (4)$$

Kokkuvõttes piisab mudelile M vastava koodi C_M saamiseks veafunktsioonist ER ning väärtusest β . Seejuures $\beta \geq 0$, sest $\beta = -\log \alpha_2$ ja $\alpha_2 \in (0, 1]$. Parameetri β väärtuse valimine on palju vaidlusi tekitanud probleem (Grünwald 1998). Mõnikord valitakse lihtsalt $\beta = 1$. Teisest küljest võib ka lähtuda hoopis parameetrist K , sest võrdusest (4) tingituna vastab igale K väärtusele kas null või üks väärtust β . Valemist (3) on näha, et K tähistab nende bittide arvu, mis kulutatakse andmete kirjeldamiseks ilma vigu arvestamata. Suhteliselt loomulik on valida K väärtuseks minimaalne mittenegatiivne täisarv, mille korral leidub võrdust (4) rahuldav väärtus β . Enne näite juurde asumist toome veel ühe definitsiooni.

Definitsioon 5 *Mudeliklassi \mathcal{M} nimetatakse heaks, kui võrdust (4) rahuldavate paaride (K, β) hulk on sama kõikide mudelite $M \in \mathcal{M}$ korral.*

Näide 5 Pöördume tagasi näite 2 juurde ning arvutame uuesti mudelite $M_1 = 01$, $M_2 = 000101$ ja $M_3 = 010101000101000101000101$ MDL-hinnangud. Seejuures jätame koodi C samaks, kuid kood C_M olgu saadud kirjeldatud meetodi abil mudelist M . Antud näite puhul saame võrdusest (4), et $K = n \cdot \log(1 + \frac{1}{2^\beta})$ (töestust vt lk 65) ning see ei sõltu vaadeldavast mudelist (järelilikult oleme valinud hea mudelite klassi. Vähim mittenegatiivne K väärtus, mille korral on see võrrand β suhtes lahenduv, on $K = 1$. Siis $1 = 32 \cdot \log(1 + \frac{1}{2^\beta})$, millest $\beta \approx 5.513$. Kasutades nüüd valemit (3) saame, et kolme vaadeldava mudeli puhul on $L_{C_{M_1}}(D) \approx 5.513 \cdot 4 + 1 = 23.052$, $L_{C_{M_2}}(D) \approx 5.513 \cdot 2 + 1 = 12.026$ ning $L_{C_{M_3}}(D) \approx 5.513 \cdot 0 + 1 = 1.000$. Mudelite M_1 , M_2 ja M_3 MDL-hinnangud tulevad siis vastavalt ligikaudu 30.052, 23.026 ja 30.000. Tulemused on väga sarnased näitega 2, kuid koodipikkused on pisut lühemad. Põhjuseks on see, et näites 2 ei olnud C_M täpne kood ning seetõttu oli võimalik lühendamine.

6. Stohhastiline keerukus

Meenutame, et MDL-i põhieesmärgiks on andmete lühike kodeerimine. Kaheosalise MDL-i puhul kodeeriti eraldi mudel ning mudeli abil andmed. Järelilikult on

ühthesid ja samu andmeid võimalik kodeerida erinevalt (erinevate mudelitega). Palju parema tulemuse saaks pakkimisel siis, kui kasutada ühte ainsat koodi, kus pole sellist liiasust. Üheks võimaluseks on kasutada stohhastilise keerukuse koodipikkusfunktsiooni $L_{SC} : E^n \rightarrow \mathbb{R}$. See funktsioon defineeritakse lähtuvalt mudelite hulgast \mathcal{M} , seega võime väita järgmist.

Stohhastiline keerukus on andmete keerukus mudelite hulga suhtes.

Kõigepealt leiame igale mudelile $M \in \mathcal{M}$ vastava koodi C_M . Kuna eesmärk on kodeerida kõik andmed võimalikult väheste bittidega, siis võiks andmed D kodeerida sellele mudelile M vastava koodiga C_M , mille puhul $L_{C_M}(D)$ on minimaalne. See vastab suurima tõepära meetodile (ik *Maximum Likelihood*) — valitakse mudel M , mille puhul andmete D tõenäosus $P(D|M)$ on suurim. Mudelit, mis andmete D jaoks minimiseerib väärtuse $L_{C_M}(D)$, tähistatakse $\widehat{M}(D)$. Niisiis kuluks andmete D kodeerimiseks

$$L_{C_{\widehat{M}(D)}}(D) = \min_{M \in \mathcal{M}} L_{C_M}(D) \quad (5)$$

bitti. Kuid MDL-i puhul on oluline see, et kõik andmeruumi kuuluvad andmed peavad saama kodeeritud ühe ja sama koodiga, mitte andmetele kõige paremini sobivale mudelile vastava koodiga. Seega võib kuluda tegelikult rohkem bitte kui $L_{C_{\widehat{M}(D)}}(D)$. Nüüd lisame koodipikkusele konstandi juurde, seda võib ette kujutada kui täiendavaid bitte, mis on vajalikud koodide ühtlustamiseks. Sellise meetodi kasutamise põhjendust vt. (Grünwald 1998).

Definitsioon 6 *Andmete D stohhastiliseks keerukuseks mudeliklassi \mathcal{M} suhtes nimetatakse väärtust*

$$L_{SC}(D) = L_{C_{\widehat{M}(D)}}(D) + K', \quad (6)$$

kus

$$K' = \log \sum_{D \in E^n} 2^{-L_{C_{\widehat{M}(D)}}(D)}. \quad (7)$$

Võrdus (7) tuleb meie soovist, et L_{SC} oleks täpne koodipikkusfunktsioon. Tõepoolest, sellisel juhul

$$1 = \sum_{D \in E^n} 2^{-L_{SC}(D)} = \sum_{D \in E^n} 2^{-L_{\widehat{M}(D)}(D) - K'} = 2^{-K'} \cdot \sum_{D \in E^n} 2^{-L_{\widehat{M}(D)}(D)},$$

millest saamegi võrduse (7). Funktsioonile L_{SC} vastavat koodi kasutamegi andmete kodeerimiseks. Koodi pikkus avaldub tänu L_{SC} definitsioonile (6) ning valemitele (3) ja (5) kujul

$$L_{SC}(D) = \min_{M \in \mathcal{M}} (\beta \cdot ER(M, D) + K) + K'$$

Juhul, kui mudeliklass \mathcal{M} on hea, s.t. samad β ja K sobivad kõikide mudelite korral, siis saame valemiks

$$L_{SC}(D) = \beta \cdot \min_{M \in \mathcal{M}} ER(M, D) + K + K'. \quad (8)$$

Tõenäosuslike mudelite korral saame (1), (5) ja (6) tõttu

$$L_{SC}(D) = \min_{P \in \mathcal{M}} (-\log P(D)) + K' = -\log \max_{P \in \mathcal{M}} P(D) + K'. \quad (9)$$

Näide 6 Siirdume jällegi näitega 2 alguse saanud uurimuse juurde. Vaatleme seekord bitistringe pikkusega $n = 4$. Tähistagu \mathcal{M}_m nende perioodiliste mudelite hulka, mille pikkus on m bitti. Leiame stohhastilised keerukused kõikvõimalikele andmetele mudeliklassi $\mathcal{M}_2 = \{00, 01, 10, 11\}$ korral. Selleks tuleb kõigepealt välja arvutada väärtused $ER(M, D)$ iga $D \in E^n$ ja iga $M \in \mathcal{M}_2$ korral (vt. tabel 1). Seejärel peab leidma iga D jaoks väärtuse $ER(\widehat{M}(D), D)$, mis on andmetel D tehtav viga selle mudeli $\widehat{M}(D) \in \mathcal{M}_2$ korral, millele andmed D kõige paremini vastavad (vt. tabeli 1 viimase veeru esimene liidetav). Valime näiteks $\beta = 1$, siis jääb veel arvutada K ja K' vastavalt valemite (4) ja (7). Seejuures võib ka kasutada spetsiaalselt mudeliklasside \mathcal{M}_m jaoks lk. 65 toodud valemiteid (10) ja (11).

7. MDL-printsiibi üheosaline versioon

MDL-printsiibi kaheosaline versioon oli mõeldud selleks, et valida mingis mudeliklassis välja parim mudel. MDL-printsiibi üheosalist versiooni rakendatakse hoopiski selleks, et otsustada, millist mudeliklassi on parem mingite andmete puhul kasutada. Näiteks võib niiviisi uurida, kas mingite andmete jaoks on paremaks mudeliks kolmanda astme polünoomid või viie peidetud neuroniga tagasisidestatud neurovõrgud.

MDL-printsiibi üheosalise versiooni järgi tuleb andmete kirjeldamiseks valida see mudeliklass, mille suhtes andmete stohhastiline keerukus on väikseim.

Tabel 1: Neljabitiste bitistringide stohhastiliste keerukuste leidmine mudeliklassi \mathcal{M}_2 suhtes $\beta = 1$ korral, stohhastiline keerukus on viimases veerus.

D	$\text{ER}(D 00)$	$\text{ER}(D 01)$	$\text{ER}(D 10)$	$\text{ER}(D 11)$	$\min_{M \in \mathcal{M}_2} \text{ER}(D M) + K + K'$
0000	0	2	2	4	$0 + 2.34 + 0.83 = 3.17$
0001	1	1	3	3	$1 + 2.34 + 0.83 = 4.17$
0010	1	3	1	3	$1 + 2.34 + 0.83 = 4.17$
0011	2	2	2	2	$2 + 2.34 + 0.83 = 5.17$
0100	1	1	3	3	$1 + 2.34 + 0.83 = 4.17$
0101	2	0	4	2	$0 + 2.34 + 0.83 = 3.17$
0110	2	2	2	2	$2 + 2.34 + 0.83 = 5.17$
0111	3	1	3	1	$1 + 2.34 + 0.83 = 4.17$
1000	1	3	1	3	$1 + 2.34 + 0.83 = 4.17$
1001	2	2	2	2	$2 + 2.34 + 0.83 = 5.17$
1010	2	4	0	2	$0 + 2.34 + 0.83 = 3.17$
1011	3	3	1	1	$1 + 2.34 + 0.83 = 4.17$
1100	2	2	2	2	$2 + 2.34 + 0.83 = 5.17$
1101	3	1	3	1	$1 + 2.34 + 0.83 = 4.17$
1110	3	3	1	1	$1 + 2.34 + 0.83 = 4.17$
1111	4	2	2	0	$0 + 2.34 + 0.83 = 3.17$

Näide 7 Proovime kõigi klasside \mathcal{M}_m jaoks välja arvutada konstantide K ja K' väärtused. Siis saab leida $D = 01010100010100010100010101010100$ stohhastilise keerukuse nende klasside suhtes ning otsustada, milline klass on D kirjeldamiseks parim. Leiame kõigepealt K , kasutades selleks valemit (4), mille järgi

$$K = \log \sum_{D \in E^n} 2^{-\beta \cdot \text{ER}(M,D)}$$

Valime suvalise mudeli $M \in \mathcal{M}_m$ ja toome sisse tähistuse

$$E_i^n = \{D \in E^n \mid \text{ER}(M, D) = i\}, \quad i = 0, 1, \dots, n.$$

Hulka E_i^n kuuluvad need andmed, millel mudel M teeb vea i . Sellega oleme jaaganud ruumi E^n lõikumatuks tükkeks, $E^n = E_0^n \dot{\cup} E_1^n \dot{\cup} \dots \dot{\cup} E_n^n$. Nüüd võime

kirjutada, et

$$\begin{aligned}
\sum_{D \in E^n} 2^{-\beta \cdot \text{ER}(M,D)} &= \sum_{D \in E_0^n} 2^{-\beta \cdot \text{ER}(M,D)} + \dots + \sum_{D \in E_n^n} 2^{-\beta \cdot \text{ER}(M,D)} = \\
&= \sum_{D \in E_0^n} 2^{-\beta \cdot 0} + \dots + \sum_{D \in E_n^n} 2^{-\beta \cdot n} = \\
&= |E_0^n| \cdot 2^{-\beta \cdot 0} + \dots + |E_n^n| \cdot 2^{-\beta \cdot n}.
\end{aligned}$$

Kuna mudel M fikseerib ära ühe „õige” n -bitise stringi (mis on saadud mudeli M korduval järjest kirjutamisel) ning vead võivad esineda ükskõik millistel i positsioonil n -st, siis on $|E_i^n| = \binom{n}{i}$. Järelikult

$$\begin{aligned}
K &= \log \sum_{i=0}^n |E_i^n| \cdot 2^{-\beta \cdot i} = \log \sum_{i=0}^n \binom{n}{i} \cdot (2^{-\beta})^i \cdot 1^{n-i} = \\
&= \log (2^{-\beta} + 1)^n = n \cdot \log \left(1 + \frac{1}{2^\beta} \right). \tag{10}
\end{aligned}$$

Huvitav on märkida, et antud juhul ei sõltu konstandi K väärtus klassist \mathcal{M}_m . Käesoleva töö autor on näidanud, et selle näite korral

$$\begin{aligned}
K' &= -K + (m(q+1) - n) \cdot \log \sum_{i=0}^q \binom{q}{i} \cdot 2^{-\beta \cdot \min(i, q-i)} + \\
&+ (n - mq) \cdot \log \sum_{i=0}^{q+1} \binom{q+1}{i} \cdot 2^{-\beta \cdot \min(i, q+1-i)}, \tag{11}
\end{aligned}$$

kus $q = \lfloor \frac{n}{m} \rfloor$, tõestus jääb mahu tõttu sellest artiklist välja. Kui K ja K' arvutada otse valemite (4), (5) ja (7), siis tuleb selleks arvutada $\text{ER}(M, D)$ iga $D \in E^n$ ja $M \in \mathcal{M}_m$ korral, mis teeb kokku $|E^n| \cdot |\mathcal{M}_m| = 2^n \cdot 2^m = 2^{n+m}$ korda. See on aga võimalik vaid väikeste n ja m korral. Kasutades valemit (11) ning selle valemi tuletamise käigus leitud algoritmi $\widehat{M}(D)$ leidmiseks on võimalik andmete D stohhastiline keerukus klassi \mathcal{M}_m suhtes välja arvutada lineaarses ajas $\mathcal{O}(n)$.

Samuti tuletas autor valemi K' leidmiseks näites 3 toodud tõenäosuslike mu-

delite klassi $\mathcal{M}_{\mathcal{P}}$ jaoks:

$$\begin{aligned}
K' = & \log \sum_{k=1}^{n-1} \frac{1}{k^k \cdot (n-1-k)^{n-1-k}} \cdot \sum_{i=0}^{\min(n-1-k, k)} \binom{n-2-k}{i-1} \cdot i^i \cdot \\
& \cdot \left[\binom{k}{i} \cdot (k-i)^{k-i} \cdot i^i \cdot (n-1-k-i)^{n-1-k-i} + \right. \\
& + \binom{k-1}{i-1} \cdot (k-i)^{k-i} \cdot (i-1)^{i-1} \cdot (n-k-i)^{n-k-i} + \\
& \left. + \binom{k-1}{i} \cdot (k-i-1)^{k-i-1} \cdot (i+1)^{i+1} \cdot (n-1-k-i)^{n-1-k-i} \right]. \tag{12}
\end{aligned}$$

Näide 8 Rakendame üheosalist MDL-i, et leida, milline klassidest $\mathcal{M}_{\mathcal{P}}$, \mathcal{M}_1 , $\mathcal{M}_2, \dots, \mathcal{M}_n$ vastab andmetele $D = 01010100010100010100010101010100$ kõige paremini. Mudeliklasside \mathcal{M}_m kohta leidsime näites 5 väärtusele $K = 1$ vastava väärtuse $\beta \approx 5.513$. Nüüd jääb veel leida K' valemist (11) ja seejärel stohhastiline keerukus $L_{SC}(D)$ valemist (8). Mudeliklassi $\mathcal{M}_{\mathcal{P}}$ korral tuleb leida K' valemist (12) ning seejärel stohhastiline keerukus valemist (9). Tulemused on ära toodud tabelis 2.

8. Kokkuvõte

Induktiivsel järeldamisel ja parima mudeli valikul on üheks võimalikuks lähemissuunaks informatsiooniteooria rakendamine. Selle vaatenurga järgi loetakse parimaks järelduseks ehk mudeliks see, mis võimaldab anda lähteandmetele lühima kirjelduse. Kolmogorovi keerukuse baasil loodud mudel võimaldab enamikel juhtudel küll parimat pakkimist, kuid paraku pole see mudel algoritmiliselt leitav. Praktilisema lahenduse pakub MDL-printsiiibi kaheosaline versioon, mille järgi on parim mudel see, mis minimiseerib mudeli kirjeldamiseks ja mudeli alusel andmete kirjeldamiseks kuluva bittide arvu summa. Paraku tuleb selleks ise valida meetod, kuidas kirjeldamine teostada (on olemas teatavad üldised nõuanded). See muudab MDL-printsiiibi rakendamise tulemuse sõltuvaks kasutaja valikutest, mis võib olla halb, sest kasutajal võib olla andmeid näinuna eelistus teatavate mudelite suhtes. Üks võimalus ühte kasutatavat koodi „automaatselt” valida on jaotises 5 kirjeldatud meetod mudelist koodi saamiseks. Samas tuleb ka sel juhul valida

Tabel 2: Andmete $D = 01010100010100010100010101010100$ stohhastiline kee-
 rukus mudeliklasside $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_n, \mathcal{M}_{\mathcal{P}}$ suhtes.

klass	β	$\min_{M \in \mathcal{M}} \text{ER}(M, D)$	K	K'	$L_{SC}(D)$	
\mathcal{M}_1	5.513	12	1.0	1.000	68.157	
\mathcal{M}_2	5.513	4	1.0	2.000	25.052	V
\mathcal{M}_3	5.513	12	1.0	3.000	70.158	
\mathcal{M}_4	5.513	4	1.0	4.000	27.052	VI
\mathcal{M}_5	5.513	12	1.0	5.000	72.156	
\mathcal{M}_6	5.513	2	1.0	5.999	18.025	I
\mathcal{M}_7	5.513	11	1.0	6.994	68.637	
\mathcal{M}_8	5.513	4	1.0	7.985	31.037	IX
\mathcal{M}_9	5.513	11	1.0	8.982	70.626	
\mathcal{M}_{10}	5.513	4	1.0	9.980	33.032	
\mathcal{M}_{11}	5.513	9	1.0	10.949	61.566	
\mathcal{M}_{12}	5.513	1	1.0	11.859	18.372	II
\mathcal{M}_{13}	5.513	10	1.0	12.769	68.900	
\mathcal{M}_{14}	5.513	4	1.0	13.680	36.731	
\mathcal{M}_{15}	5.513	12	1.0	14.590	81.746	
\mathcal{M}_{16}	5.513	4	1.0	15.500	38.552	
\mathcal{M}_{17}	5.513	11	1.0	16.468	78.111	
\mathcal{M}_{18}	5.513	2	1.0	17.436	29.462	VII
\mathcal{M}_{19}	5.513	10	1.0	18.406	74.536	
\mathcal{M}_{20}	5.513	2	1.0	19.375	31.401	X
\mathcal{M}_{21}	5.513	9	1.0	20.344	70.960	
\mathcal{M}_{22}	5.513	2	1.0	21.313	33.339	
\mathcal{M}_{23}	5.513	7	1.0	22.281	61.872	
\mathcal{M}_{24}	5.513	0	1.0	23.249	24.249	IV
\mathcal{M}_{25}	5.513	6	1.0	24.219	58.296	
\mathcal{M}_{26}	5.513	1	1.0	25.188	31.701	XI
\mathcal{M}_{27}	5.513	4	1.0	26.156	49.208	
\mathcal{M}_{28}	5.513	1	1.0	27.126	33.638	
\mathcal{M}_{29}	5.513	2	1.0	28.094	40.119	
\mathcal{M}_{30}	5.513	1	1.0	29.062	35.575	
\mathcal{M}_{31}	5.513	0	1.0	30.031	31.031	VIII
\mathcal{M}_{32}	5.513	0	1.0	31.000	32.000	XII
$\mathcal{M}_{\mathcal{P}}$	—	—	—	4.574	23.614	III

kasutajal vähemalt ühe parameetri (kas K või β) väärtus. Käesolevas töös on valitud minimaalne täisarvuline K , millele leidub vastav β . Võimalusi on ka muid, vt. näiteks (Grünwald 1998).

Kuna kaheosaline versioon MDL-ist kasutab kahte eraldi koodi andmete ja mudelite jaoks, siis on tulemus kaugel optimaalsest. Üheosaline MDL kasutab kodeerimisel ühtset, stohhastilise keerukuse koodi. Nõnda tekib olukord, kus kodeerimine ei toimu enam kasutades üht mudelit, vaid tervet mudelite klassi. Üheosalist MDL-i kasutatakse seepärast mudeliklasside võrdlemiseks. Näitest selgub, et stohhastilise keerukuse leidmine ei ole triviaalne ülesanne, seepärast kasutatakse üheosalisele MDL-ile mitmesuguseid lähendeid (Baxter & Oliver 1994), mis on omaette uurimisteema ja jäi siinsest tööst välja.

Näitena on käesolevas töös uuritud bitistringide perioodilisi ja tõenäosuslikke mudeleid. Leitud on seos parameetrite β ja K vahel, mis võimaldab hõlpsasti rakendada MDL-printsipi suvaliste andmete korral. Selle näite kohta on leitud stohhastilise keerukuse arvutamist lihtsustavad valemid, need on esitatud tõestuseta. Tulevikus võiks selle näite kohta teha põhjalikemaid võrdlusi erinevate meetodite vahel, uurida võiks parameetrite β ja K väärtuste valiku mõju tulemustele.

Viited

- Baxter, R. A., and Oliver, J. J. 1994. MDL and MML: similarities and differences. Technical report.
- Domingos, P. 1998. Occam's two razors: The sharp and the blunt. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press.
- Grünwald, P. D. 1998. *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph.D. Dissertation, University of Amsterdam. Available as ILLC Dissertation Series 1998-03; see <http://homepages.cwi.nl/pdg/thesispage.html>.
- Hansen, M. H., and Yu, B. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96(454):746–774.
- Kiho, J. 2003. *Algoritmid ja andmestruktuurid*. Tartu Ülikooli Kirjastus.
- Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.