

Logiikan uusimpia suuntauksia

6. Lisää selitettävyydestä

Lähteen [1] Lause 4.0.2

Tutustutaan sitten lähteen [1] lauseeseen 4.0.2 ja tätä edeltävään propositioon 4.0.1. Lauseen 4.0.2 merkitys tutkimusartikkelin [2] tulosten kannalta on seuraava:

- ▶ Osoittamalla, että erityisen selitysongelman tulosteet saadaan yksinkertaisesti kaavojen φ alikonjunktioista χ (ks. lauseen väite), tarkoittaa, että potentiaalisia tulosteita on huomattavasti vähemmän, kuin aluksi voisi olettaa.
- ▶ Täten voidaan luoda algoritmi, joka etsii tulosteita (eli selityksiä) tehokkaasti, koska väärää muotoa olevia mahdollisia tulosteita ei tarvitse lainkaan tarkistaa.
- ▶ *Tehokkuus* tässä tarkoittaa, että etsintä voidaan käytännössä tehdä menestyksekkäästi. Ilman lauseen 4.0.2 tulosta, tulosteita ei käytännössä voisi menestyksekkäästi etsiä.

[1] Ilpo Lehtimäki. Ilpo Lehtimäki. *Selitettävyydestä propositiologiikassa*. Tampereen yliopisto, 2023.

[2] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood F. Rankooh, and Miikka Vilander. *Explainability via short formulas: the case of propositional logic with implementation*. In RCRA 2022, CEUR Workshop Proceedings, pages 64–77. CEUR-WS.org, 2022.

Selitettävyyys ja käytännön data

Lähde [1] perustui artikkeliin [2]. Tutustutaan seuraavaksi lähteen [3] **joihinkin** tuloksiin. Lähde [3] **ei tarvitse** suoraan lukea, mutta kuten muutkin lähteet, se on saatavissa verkosta; suoraviivaisinta lienee tarkastella seuraavaa versiota:

<https://arxiv.org/pdf/2307.06971.pdf>

[3] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Feyzbakhsh Rankooh, Miikka Vilander: Short Boolean Formulas as Explanations in Practice. *European Conference on Logics in Artificial Intelligence (JELIA)* 2023.

Selitettävyyys ja käytännön data

Palautetaan mieleen **unaarimallit** osiosta 1. Kyseessä on siis mallit, joiden aakkosto on $\tau = \{P_1, \dots, P_n\}$, missä $P_1, \dots, P_n$ ovat yksipaikkaisia predikaattisymboleja.

Palautetaan mieleen myös **propositionaalisen τ -tyypin** käsite. Muistamme, että τ -mallin domainin M voi **osittaa** propositionaalisten τ -tyyppien avulla: osiossa 1 havaitsimme, että jokaiselle unaariselle τ -mallille $\mathfrak{M}$ on olemassa jokin joukko $\mathcal{U}_{\mathfrak{M}}$ propositionaalisia τ -tyyppejä, joiden tulkinnat osittavat domainin M (ks. osio 1). Tällöin sanomme, että $\mathcal{U}_{\mathfrak{M}}$ **luo osituksen malliin $\mathfrak{M}$** .

Selitettävyyys ja käytännön data

Propositionaalinen τ -tyyppi määriteltiin osajoukoksi

$$T \subseteq \tau = \{P_1, \dots, P_n\}.$$

Usein kirjallisuudessa propositionaalinen τ -tyyppi T kuitenkin määritellään hieman eri tavalla, joka on kuitenkin täysin ekvivalentti esittämämme tavan kanssa. Tarkastellaan tätä seuraavaksi.

Aluksi, aakkosto $\tau = \{P_1, \dots, P_n\}$ korvataan aakkostolla $\Pi = \{p_1, \dots, p_n\}$. Tässä siis yksinkertaisesti yksipaikkaiset predikaattisymbolit P_k korvataan propositiosymboleilla p_k .

Tarkastellaan propositionaalista τ -tyyppiä $T \subseteq \tau$. Seuraavaa konjunktia kutsutaan tyyppin T **propositiologiikan vastinekaavaksi**:

$$\bigwedge_{P_i \in T} p_i \quad \wedge \quad \bigwedge_{P_i \in \tau \setminus T} \neg p_i$$

Selitettävyyys ja käytännön data

Huomaa, että tarkkaan ottaen kaava

$$\bigwedge_{p_i \in T} p_i \quad \wedge \quad \bigwedge_{p_i \in \tau \setminus T} \neg p_i$$

voidaan määritellä monella eri tavalla, esim. vaihtelemalla sulutusta ja propositiosymbolien järjestystä. Oletamme kuitenkin, että näille kaavoille on kiinnitetty joku standardisulutus ja propositioiden järjestys, jolloin siis τ -tyypillä T on täsmälleen yksi propositiologiikan vastinekaava.

Selitettävyys ja käytännön data

Esimerkki: Oletetaan, että $\tau = \{P_1, P_2\}$. Tällöin τ -tyypin $\{P_1\}$ vastinekaava on $p_1 \wedge \neg p_2$. Samaan tapaan, τ -tyypin $\{P_1, P_2\}$ vastinekaava on $p_1 \wedge p_2$, ja τ -tyypin $\emptyset$ vastinekaava on $\neg p_1 \wedge \neg p_2$.

Havaitsemme, että vastinekaavat ovat maksimaalisia konjunktioita suhteessa aakkostoon Π . Tyypin T vastinekaavan **tulkinta** (**ekstensio**, **totuusjoukko**) mallissa $\mathfrak{M}$ määritellään samaksi, kuin tyypin T tulkinta (ekstensio, totuusjoukko).

Koska τ -tyyppi ja sen vastinekaava φ_τ käyttäytyvät oleellisesti samaan tapaan, voimme kutsua vastinekaavaa yksinkertaisesti τ -tyypiksi. Voimme myös samastaa aakkostot $\tau = \{P_1, \dots, P_n\}$ ja $\Pi = \{p_1, \dots, p_n\}$. Käytännön kannalta tämä samastus on hyödyllinen. Formaalisakaan mielessä se ei aiheuta minkäänlaista vaaraa, koska asiayhteydestä on tyypillisesti riittävän selvää, mitä tarkoitetaan. Jatkossa siis voimme puhua τ -tyypeistä ja tarkoittaa sillä τ -tyypin vastinekaavaa.

Selitettävyyys ja käytännön data

Määritellään seuraava käännös f propositiologiikan Π -kaavoilta predikaattilogiikan τ -kaavoille $\psi(x)$. Edelleen siis $\tau = \{P_1, \dots, P_n\}$ ja $\Pi = \{p_1, \dots, p_n\}$.

$$\begin{aligned}f(p_i) &= P_i(x) \\f(\neg\varphi) &= \neg f(\varphi) \\f(\varphi \wedge \varphi') &= f(\varphi) \wedge f(\varphi') \\f(\varphi \vee \varphi') &= f(\varphi) \vee f(\varphi')\end{aligned}$$

Propositiologiikan kaavan φ **tulkinta** (**extensio**, **totuusjoukko**) predikaattilogiikan τ -mallissa $\mathfrak{M}$ määritellään kaavan $f(\varphi)$ määräämäksi unaarirelaatioksi, eli relaatioksi

$$\{a \in M \mid \mathfrak{M}, \{(x, a)\} \models f(\varphi)\}$$

Mikäli φ on propositiologiikan kaava, voimme merkitä kaavan φ tulkintaa kirjoittamalla $\varphi^{\mathfrak{M}}$. Näin olemme siis onnistuneet tulkitsemaan propositiologiikan predikaattilogiikan τ -mallissa unaarisella aakkostolla τ . Tulkinnat $\varphi^{\mathfrak{M}}$ ovat yksipaikkaisia relaatioita joukossa M .

Selitettävyyys ja käytännön data

Huomautettakoon vielä, että τ -mallin piste $u \in M$ vastaa luonnollisesti valuaatiota v_u aakkostolle Π siten, että $v_u(p_i) = 1$ jos ja vain jos $u \in P_i^{\mathfrak{M}}$. On siis luonnollista ajatella mallia $\mathfrak{M}$ vaihtoehtoisesti joko unaarimallina aakkostolle τ , tai kokoelmana Π -valuaatioita.

Huom., kyseinen kokoelma Π -valuaatioita sallii useamman esiintymän samalle valuaatiolle. Eli vaikkapa se Π -valuaatio, joka tulkitsee jokaisen symbolin $p_i \in \Pi$ todeksi, voi esiintyä *monta* kertaa kyseisessä kokoelmassa. Tekninen termi tällaiselle kokoelmalle on **multijoukko**. Multijoukko on pari (S, m) , jossa S on joukko ja $m : S \rightarrow \mathbb{Z}_+$ kertoo jokaisen alkion $u \in S$ esiintymien lukumäärän $m(u)$. (Tässä siis $\mathbb{Z}_+$ tarkoittaa positiivisia kokonaislukuja.) Joskus multijoukkoja kirjoitetaan muotoon $\{a, a, a, b, b\}$, missä siis esiintyy alkio a kolme ja alkio b kaksi kertaa.

Unaarimallin $\mathfrak{M}$ voi siis vaihtoehtoisesti tulkita multijoukoksi Π -valuaatioita. Formaalisti kyseessä on tietenkin kaksi eri oliota, unaarimalli ja multijoukko valuaatioita. Käytännön kannalta eroa ei kuitankaan ole. Kontekstista selviää aina, onko tarpeen kiinnittää huomiota kyseiseen formaaliin eroon vai ei.

Selitettävyyys ja käytännön data

Propositiologiikan kaavan φ **todennäköisyys** mallissa $\mathfrak{M}$ saadaan luonnolliseen tapaan kaavasta

$$\Pr(\varphi) = \frac{|\varphi^{\mathfrak{M}}|}{|M|}.$$

Tässä siis φ on Π -kaava ja $\mathfrak{M}$ on τ -malli. Mutta, kuten yllä todettu, ero aakkostojen Π ja τ välillä on lähinnä muodollinen, joten käytännössä voimme puhua myös propositiologiikan τ -kaavoista ja tarkastella aakkostoa τ propositiologiikan aakkostona.

Selitettävyyys ja käytännön data

Määritellään seuraavaksi luonnollinen yleistys yleiselle selitysongelmalla.

Tarkastellaan logiikoita

$$\blacktriangleright L_1 = (\mathcal{M}, \mathcal{F}_1, v_1, m_1),$$

$$\blacktriangleright L_2 = (\mathcal{M}, \mathcal{F}_2, v_2, m_2).$$

Logiikoiden mallijoukko on siis sama. Oletetaan lisäksi, että semantiikkafunktiot v_1 ja v_2 ovat tyyppiä

$$\blacktriangleright v_1 : \mathcal{M} \times \mathcal{F}_1 \rightarrow V,$$

$$\blacktriangleright v_2 : \mathcal{M} \times \mathcal{F}_2 \rightarrow V.$$

Funktioilla v_1 ja v_2 on siis sama maalijoukko V , eli molemmat logiikat L_1 ja L_2 ovat V -arvoisia.

Oletetaan sitten, että olemme kiinnittäneet **metriikan** d joukolle V , eli siis oletetaan, että d on funktio $d : V \times V \rightarrow \mathbb{R}_{\geq 0}$, jolle pätee (vertaa määritelmä ?? näissä dioissa)

$$1. \quad d(x, y) = 0 \Leftrightarrow x = y$$

$$2. \quad d(x, y) = d(y, x)$$

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

Tässä siis $\mathbb{R}_{\geq 0}$ on joukko $\{r \in \mathbb{R} \mid r \geq 0\}$.

Selitettävyyys ja käytännön data

Yleisesti metriikat kuvaavat abstrakteja etäisyyskäsitteitä. Yllä määrittelemämme metriikka $d : V \times V \rightarrow \mathbb{R}_{\geq 0}$ mahdollistaa logiikkojen L_1 ja L_2 kaavojen ekvivalenttisuuden asteen tarkastelun seuraavaan tapaan.

Olkoon φ logiikan L_1 ja ψ logiikan L_2 kaavoja. Merkitään $\varphi \equiv_r \psi$ jos $d(v_1(\varphi), v_2(\psi)) \leq r$. Intuitiivisesti tämä tarkoittaa, että kaavat φ ja ψ eroavat toisistaan korkeintaan r yksikköä. Määritellään relaatio $\equiv_r$ symmetriseksi, eli sallitaan myös merkintä $\psi \equiv_r \varphi$ jos $\varphi \equiv_r \psi$.

Metriikan d lisäksi, oletetaan, että olemme määritelleet **reaalilukujen esityksen**, eli jonkin funktion $r : \mathcal{R} \rightarrow \mathbb{R}$, jossa $\mathcal{R}$ on yksinkertaisesti jokin äärellinen tai numeroituvasti ääretön symbolijoukko. Intuitiivisesti $\mathcal{R}$ tosiaan mahdollistaa reaalityyppien käyttämisen syötteinä tietokoneellistettaville algoritmeille. $\mathcal{R}$ voisi siis olla vaikkapa jokin äärellinen kokoelma desimaaliesitysmuodossa annettuja rationaalilukujen esityksiä (esimerkiksi symbolijono `0.78` voisi olla $\mathcal{R}$:n alkio), jotka funktio r tulkitsee vastaaviksi luvuiksi (eli jos syötteenä funktiolle r on symbolijono `0.78` $\in \mathcal{R}$, niin tulosteena on reaaliluku `0.78` $\in \mathbb{R}$). Myös muunlaiset ratkaisut ovat mahdollisia. Oleellista on, että $\mathcal{R}$ on numeroituva symbolijoukko ja $r : \mathcal{R} \rightarrow \mathbb{R}$ funktio.

Selitettävyyys ja käytännön data

Määritelmä 5.1

Olkoot $L_1 = (\mathcal{M}, \mathcal{F}_1, v_1, m_1)$ ja $L_2 = (\mathcal{M}, \mathcal{F}_2, v_2, m_2)$ eri logiikoita ja $v_1 : \mathcal{M} \times \mathcal{F}_1 \rightarrow V$ ja $v_2 : \mathcal{M} \times \mathcal{F}_2 \rightarrow V$ niiden semantiikkafunktiot.

Logiikan L_1 yleinen approksimatiivinen selitysongelma logiikalla L_2 määritellään seuraavasti.

1. Syöte: (φ, ϵ) missä φ on logiikan L_1 kaava ja $\epsilon \in \mathcal{R}$.
2. Tuloste: logiikan L_2 kaava ψ jonka arvo $m(\psi)$ on pienin mahdollinen niiden logiikan L_2 kaavojen χ joukossa, joille pätee $\chi \equiv_{r(\epsilon)} \varphi$.

Intuitiivisesti saamme siis tulosteeksi lyhyen kaavan ψ , joka eroaa korkeintaan lukua $r(\epsilon)$ vastaavan “samankaltaisuusasteen” verran kaavasta φ .

Selitettävyyys ja käytännön data

Määritelmä 5.2

Olkoot $L_1 = (\mathcal{M}, \mathcal{F}_1, v_1, m_1)$ ja $L_2 = (\mathcal{M}, \mathcal{F}_2, v_2, m_2)$ eri logiikoita ja $v_1 : \mathcal{M} \times \mathcal{F}_1 \rightarrow V$ ja $v_2 : \mathcal{M} \times \mathcal{F}_2 \rightarrow V$ niiden semantiikkafunktiot. **Logiikan L_1 yleinen vastineongelma logiikalla L_2 määritellään seuraavasti.**

1. Syöte: logiikan L_1 kaava φ .
2. Tuloste: logiikan L_2 kaava ψ jonka arvo $d(v_1(\varphi), v_2(\psi))$ on pienin mahdollinen.

Intuitiivisesti saamme siis tulosteen logiikan L_2 kaavan, joka vastaa alkuperäistä kaavaa φ . (Tulostekaava ψ ei tässä tapauksessa ole välttämättä lyhyt, joten seuraavana kysymyksenä voisi olla etsiä mahdollisimman lyhyt kaava ψ' , joka kelpaa tulostekaavaksi.)

Selitettävyyys ja käytännön data

Sovelletaan määritelmän 5.2 kuvaamaa vastineongelmaa käytännön datan tutkimiseen.

Tarkastellaan τ^+ -mallia $\mathfrak{M}$ jonka aakkosto τ^+ on äärellinen joukko yksipaikkaisia predikaatteja. Kuten havaitsimme yllä, malli $\mathfrak{M}$ vastaa multijoukkoa propositiologiikan valuaatioita. Olkoot siis

- ▶ $\tau = \{p_1, \dots, p_n\}$ ja
- ▶ $\tau^+ = \{p_1, \dots, p_n\} \cup \{q\}$

propositiologiikan aakkostoja. Nyt siis $\mathfrak{M}$ on τ^+ -malli, joten tässä kohtaa teimme aikaisemmin yllä kuvaillun samastuksen predikaattilogiikan ja propositiologiikan aakkostojen kanssa.

Selitettävyyys ja käytännön data

Olkoon logiikkana L_1 propositiologiikka aakkostolla $\{q\}$ ja logiikkana L_2 propositiologiikka aakkostolla $\{p_1, \dots, p_n\}$. Tavoitteenamme on selittää kaava q mahdollisimman lyhyillä, logiikan L_2 kaavoilla. Tämä vastaa seuraavaa käytännön tilannetta.

1. $\mathfrak{M}$ on joukko dataa, esimerkiksi jokin sopiva tietokanta.
2. q on jokin tärkeä ominaisuus, jonka haluamme ymmärtää tarkemmin. Jos mallin $\mathfrak{M}$ alkiot olisivat vaikkapa jonkin pankin lainanhakijoita, $q^{\mathfrak{M}} \subseteq M$ voisi olla niiden asiakkaiden joukko jotka saivat lainaa.
3. Propositiot $p_1, \dots, p_n$ vastaavat muita asiakkaiden tärkeitä ominaisuuksia, vaikkapa “henkilön tilin saldo on positiivinen”, “henkilö on hakenut mediaanikeston ylittävää lainaa”, “henkilö työskentelee johtotehtävissä”, jne.

Tavoitteenamme on siis selittää q (lainan saanti) propositioiden $p_1, \dots, p_n$ avulla. Koska q ja $p_1, \dots, p_n$ voidaan samastaa myös yksipaikkaisten predikaattien (ja täten yksipaikkaisten relaatioiden) kanssa, on luonnollista puhua *attribuuteista* $q, p_1, \dots, p_n$. Tällöin q on *kohdeattribuutti*.

Selitettävyyys ja käytännön data

Palautetaan mieleen, että $S \oplus U$ tarkoittaa joukkojen S ja U symmetristä erotusta, eli joukkoa

$$(S \setminus U) \cup (U \setminus S)$$

joka siis sisältää pisteet, jotka kuuluvat täsmälleen yhteen joukoista U ja S .

Valitaan virheen käsitteeksi symmetrisen erotuksen todennäköisyyttä vastaava luku, eli τ -kaavan ψ **virhe suhteessa kaavaan** q on

$$\frac{|q^{\mathfrak{M}} \oplus \psi^{\mathfrak{M}}|}{|M|}$$

Huomaa, että

$$q^{\mathfrak{M}} \oplus \psi^{\mathfrak{M}} = (\neg(q \leftrightarrow \psi))^{\mathfrak{M}}$$

eli kaavan ψ virhe suhteessa kaavaan q antaa niiden pisteiden osuuden, jotka ovat “eri mieltä” ψ :n ja q :n totuusarvoista.

Selitettävyyys ja käytännön data

Tarkastellaan sitten logiikan L_2 fragmentteja, eli osalogiikoita. Olkoon s positiivinen kokonaisluku. Merkitköön L_2^s logiikkaa, joka on muuten täsmälleen sama kuin L_2 , mutta kaavoiksi hyväksytään ainoastaan kaavat, joiden pituus on korkeintaan s .

Voimme nyt etsiä eriasteisia selityksiä q :lle seuraavaan tapaan. Kutsumme prosessia **iteroidusti tarkentavaksi selitysalgoritmiksi**.

1. Etsitään aluksi ratkaisu logiikan L_1 yleiseen vastineongelmaan logiikalla L_2^s kun syötteenä on q ja $s = 1$. Eli etsitään q :lle selitys, jonka pituus on korkeintaan s
2. Toistetaan tämä arvolla $s = 2$.
3. Jatketaan tätä aina yhtä isommalla s :n arvolla, kunnes parempaa kaavaa ei enää voi saavuttaa. Huomaa, että yläraja kaavan pituudelle on olemassa seuraavalla sivulla olevan proposition nojalla.

Selitettävyyden ja käytännön data

Propositio 5.3

Olkon $\tau = \{p_1, \dots, p_n\}$. On olemassa τ -kaava ψ siten, että kaikille τ -kaavoille ψ' , kaavan ψ virhe suhteessa kaavaan q on pienempi tai yhtäsuuri kuin kaavan ψ' virhe suhteessa kaavaan q .

Todistus. Tässä riittää havaita, että propositiologiikan τ -valuaatioita on $2^{|\tau|}$ kappaletta. Täten τ -ominaisuuksia (eli joukkoja τ -valuaatioita) on $2^{2^{|\tau|}}$ kappaletta. Aikaisemmin kurssilla olemme todistaneet, että kaikki τ -ominaisuudet ovat määriteltävissä propositiologiikassa. Täten voimme jakaa kaikki τ -kaavat yhteensä $2^{2^{|\tau|}}$:n ekvivalenssiluokkaan, joista jokaisessa luokassa on vain keskenään ekvivalentteja kaavoja. Voimme lisäksi valita jokaiselle luokalle edustajan, joka on luonnollisesti jotakin äärellistä pituutta. Nyt siis voimme ottaa maksimin näistä pituuksista ja käydä läpi kaikki kaavat siihen pituuteen asti. $\square$

Selitettävyys ja käytännön data

Todistus 2. Osiossa “2. Logiikoista” olevan Proposition 1.13 todistuksen nojalla jokainen τ -ominaisuus voidaan määritellä kaavalla

$$\bigvee_{v_i \in \mathcal{O}} \varphi_{v_i}$$

missä φ_{v_i} on konjunktio $r_i \wedge \dots \wedge r_n$ jossa

- ▶ $r_\ell = p_\ell$ jos $v_i(p_\ell) = 1$ ja
- ▶ $r_\ell = \neg p_\ell$ jos $v_i(p_\ell) = 0$.

Havaitsemme, että φ_{v_i} on yksinkertaisesti τ -tyyppi. Näin ollen jokainen ominaisuus on disjunktio tyypeistä. (Huomautettakoon, että disjunktio tyhjistä joukosta on määritelmän mukaan ristiriita, joka voidaan ilmaista kaavalla $p_1 \wedge \neg p_1$.) jatkuu...

Selitettävyyys ja käytännön data

Näin ollen siis *jokainen τ -kaava on ekvivalentti jonkin kaavan kanssa, joka on disjunktio τ -tyypeistä* (mukaanluettuna mahdollisuutena tyhjä disjunktio, joka koodataan kaavalla $p_1 \wedge \neg p_1$).

Tehdään nyt yksinkertaisesti seuraavasti. Käydään läpi kaikki τ -tyypit t ja määritellään τ -tyypeistä koostuva joukko S seuraavasti.

- ▶ Oletetaan, että t toteutuu τ^+ -mallissa $\mathfrak{M}$, mikä tarkoittaa, että $t^{\mathfrak{M}} \neq \emptyset$. Tällöin
 - ▶ $t \in S$ jos $|(t \wedge q)^{\mathfrak{M}}| \geq |(t \wedge \neg q)^{\mathfrak{M}}|$
 - ▶ muutoin $t \notin S$.
- ▶ Oletetaan, että t ei toteudu mallissa $\mathfrak{M}$, eli $t^{\mathfrak{M}} = \emptyset$. Tällöin $t \notin S$.

Tässä siis lisäsimme joukkoon S ne tyypit, joiden tulkinnassa $t^{\mathfrak{M}}$ olevista alkioista vähintään puolet toteuttaa q :n (eli kuuluvat joukkoon $q^{\mathfrak{M}}$). Jätimme pois ne tyypit joiden tulkinnan alkioista alle puolet toteuttavat q :n. Jatkuu...

Selitettävyyys ja käytännön data

Tarkastellaan kaavaa $\sqrt{S}$. On helppo nähdä, että kaikista τ -kaavoista, τ -kaavan $\sqrt{S}$ virhe suhteessa kaavaan q on pienin mahdollinen. Täten siis kaavan $\sqrt{S}$ virhe on minimaalinen. $\square$

Selityttävyys ja käytännön data

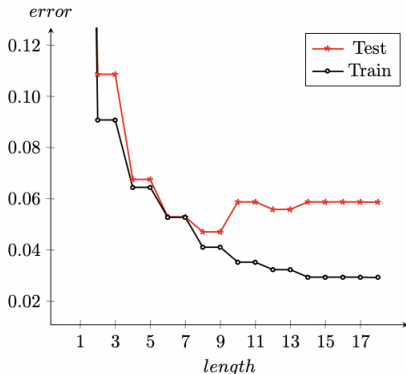
Todistuksen kaavaa $\forall S$ kutsutaan **ideaaliluokittelijaksi** mallille $\mathfrak{M}$. Ideaaliluokittelija on paras mahdollinen approksimaatio kaavalle q suhteessa malliin $\mathfrak{M}$.

Selitettävyyys ja käytännön data

Iteroidusti tarkentavasssa selitysalgoritmissa ei kuitenkaan käytännössä kannata jatkaa siihen asti, että saadaan ideaaliluokittelija (tai sen kanssa ekvivalentti kaava) tulosteena. Miksi?

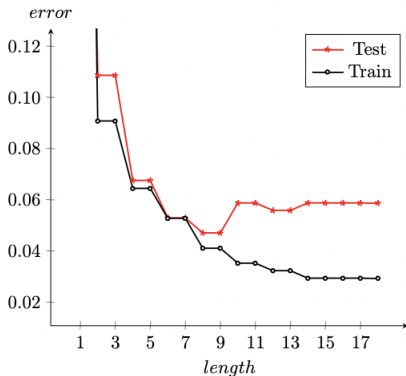
Kyseessä on nk. *ylisovittaminen* (englanniksi overfitting). Käytännössä malli $\mathcal{M}$ on lähes aina liian pieni datajoukko vastataksaan todellisuutta riittävän tarkasti (liian pieni otos). Kun aakkostona on $\tau = \{p_1, \dots, p_n\}$, on tyyppejä $2^{|\tau|}$, joka voi jo pienellä n olla erittäin suuri. Usein tarvittaisiinkin astronomisen suuri mallin $\mathcal{M}$ domaini, jotta malli $\mathcal{M}$ vastaisi todellisuutta niin tarkasti, että jopa mallista saatu ideaaliluokittelija vastaisi todellisuutta. Katsotaan tähän liittyviä esimerkkejä lähteestä [3].

Cancer data set Wisconsin



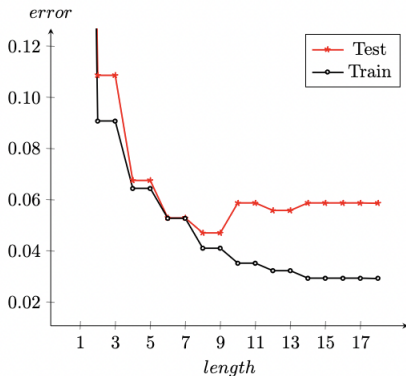
Kuvassa iteroidusti tarkentavan selitysalgoritmin tuloksia. Propositio q kysyy, “onko kasvain hyvänlaatuinen”. Propositioita $p_1, \dots, p_n$ on 9 kappaletta ja alkuperäisen mallin $\mathfrak{M}$ (eli datajoukon) koko on 683 pistettä.

Cancer data set Wisconsin



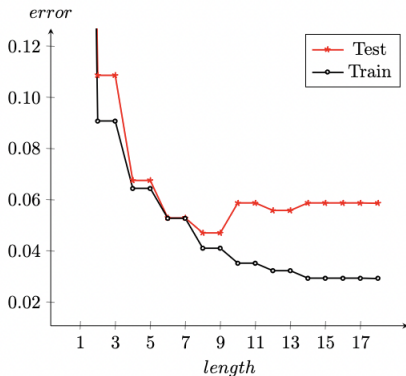
Malli $\mathcal{M}$ on jaettu kahtia nk. training-malliin $\mathcal{M}_{\text{training}}$ (342 pistettä) ja testing-malliin $\mathcal{M}_{\text{test}}$ (341 pistettä), eli $\mathcal{M}$ on näiden kahden mallin erillinen yhdiste. Iteroidusti tarkentavaa selitysalgoritmia ajetaan mallilla $\mathcal{M}_{\text{training}}$. Tästä saadaan musta (eli alempi) käyrä, joka antaa parhaimman selityskaavan virheen suhteessa kaavaan q mallissa $\mathcal{M}_{\text{training}}$. Vaaka-akselina on parhaan selityskaavan sallittu maksimipituus, jota siis kasvatetaan.

Cancer data set Wisconsin



Saamme siis mallista $\mathcal{M}_{\text{training}}$ parempia ja parempia (eli virhe pienempi) selityskaavoja ψ_s kun maksimipituus s kasvaa. Punainen käyrä saadaan laskemalla näiden kaavojen ψ_s virhe suhteessa kaavaan q mallissa $\mathcal{M}_{\text{test}}$. Kaavat siis saadaan mallista $\mathcal{M}_{\text{training}}$, mutta tämän jälkeen ne vielä testataan mallissa $\mathcal{M}_{\text{test}}$. Noin pituudesta 9 eteenpäin, käyrät alkavat kehittymään selvästi eri suuntiin. Tämä (9) on ylisovittumispiste.

Cancer data set Wisconsin



Ylisovittuminen siis tarkoittaa, että nyt kaavat ψ_s ovat jo liian tarkkoja ja puhuvat enemmänkin datamallista $\mathcal{M}_{\text{training}}$ kuin mistään yleisestä ilmiöstä. Datajoukot eivät yksinkertaisesti ole tarpeeksi suuria, jotta tarkempaa kaavaa kannattaisi laskea. (Ylisovittumispisteen löytämiselle ei ole olemassa mitään yksiselitteisesti parasta teoriaa, vaan se tehdään tapauskohtaisesti. Tässä tapauksessa se tehdään yksinkertaisesti kuvaajan avulla.)

Cancer data Wisconsin

Pituutta 9 tai alle olevat kaavat kelpaavat q :n selityksiksi. Valitaan pituutta 8 oleva kaava, sillä sen virhe on sama, kuin pituutta 9 olevan kaavan. q siis kysyy, “onko kasvain hyvänlaatuinen“. Selityskaava näyttää seuraavalta:

$$\neg(((p \wedge u) \vee r) \wedge s)$$

p = thick clump

u = bare nuclei

r = single epithelial cell size

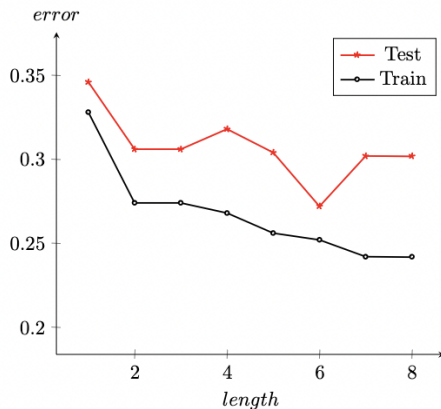
s = uniform cell shape

Kaavan virhe mallissa $\mathfrak{M}_{training}$ on 0.047. Esim. neuroverkot antavat virheen 0.032 lähteessä [4]. Lyhyen Boolean kaavan virhe on vain hieman suurempi. Toisaalta, neuroverkkoa on suoraan lähes mahdoton tulkita, kun taas yllä oleva lyhyt kaava on alan ekspertille täysin yksinkertainen ja selittää propositiota q .

Selitettävyys ja käytännön data

Katsotaan sitten toista datajoukkoa. Kaikki periaatteet ovat samat kuin cancer data -esimerkissä. Nyt q kysyy “kuka saa lainan”.

Statlog—German credit data set: Kuka saa lainan?



1000 datapistettä, 68 propositiota

Statlog—German credit data set: Kuka saa lainan?

Selitykseksi saadaan todella lyhyt kaava $\neg(b \wedge a) \vee e$, eli

$$\neg(\text{negative_balance} \wedge \text{above_median_loan_duration}) \\ \vee \text{employment_on_managerial_level}$$

Pituus: 6

Virhe: 0.27

Neuroverkoilla saadaan virhe 0.24 artikkelissa [5]. Nyt siis Boolean kaavan selitysvaima on jo radikaalin hyvä, koska virhekin on käytännössä sama kuin vastaavalla neuroverkolla.

[5] Josephine Griffith, Paul O'Dea, and Colm O'Riordan. A neural net approach to data mining: Classification of users to aid information management. In Piotr S. Szczepaniak, Javier Segovia, Janusz Kacprzyk, and Lotfi A. Zadeh, editors, Intelligent Exploration of the Web, pages 389-401. Physica-Verlag HD, Heidelberg, 2003.

Lähteet

- [1] Ilpo Lehtimäki. Ilpo Lehtimäki. *Selitetävyydestä propositiologiikassa*. Tampereen yliopisto, 2023.
- [2] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood F. Rankooh, and Miikka Vilander. *Explainability via short formulas: the case of propositional logic with implementation*. In RCRA 2022, CEUR Workshop Proceedings, pages 64–77. CEUR-WS.org, 2022.
- [3] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Feyzbakhsh Rankooh, Miikka Vilander: Short Boolean Formulas as Explanations in Practice. *European Conference on Logics in Artificial Intelligence (JELIA)* 2023.
- [4] Branko Ster and Andrej Dobnikar. Neural networks in medical diagnosis: Comparison with other methods. In Proceedings of the International Conference on Engineering Applications of Neural Networks, 1996.
- [5] Josephine Griffith, Paul O’Dea, and Colm O’Riordan. A neural net approach to data mining: Classification of users to aid information management. In Piotr S. Szczepaniak, Javier Segovia, Janusz Kacprzyk, and Lotfi A. Zadeh, editors, *Intelligent Exploration of the Web*, pages 389–401. Physica-Verlag HD, Heidelberg, 2003.