Iso data kuriin ja järjestykseen

Markus Hotakainen

Yksi Suomen Akatemian huippuyksiköistä on ”Suomalainen laskennallisen päättelyn huippuyksikkö”, jota johtaa professori Samuel Kaski Aalto-yliopistosta. Tutkimuksen tavoitteena on kehittää ja muokata menetelmiä, joilla suuria datamääriä voidaan muuntaa hyödylliseksi informaatioksi.

Huippuyksikön työssä keskeisiä käsitteitä ovat ”big data” ja koneoppiminen. Mutta mitä laskennallinen päättely käytännössä tarkoittaa?

– Laskennallisella päättelyllä on toki moniakin merkityksiä, mutta meillä se tarkoittaa erityisesti kahta toisiinsa liittyvää asiaa. Toinen niistä on tilastollinen päättely eli aineiston perusteella johdetaan malli ja sen pohjalta laaditaan ennusteita. Toinen on laskennallinen logiikka eli tietyillä reunaehdoilla tehdään johtopäätöksiä, listaa Samuel Kaski.

Laskennallisessa päättelyssä tehdään sekä aitoja ennusteita, mutta tavallaan myös taannehtivia ennusteita tulevaisuudesta: katsotaan, saadaanko datan pohjalta tulos, joka tiedetään jo entuudestaan.

– Koneoppimisessa on keskeistä, että aineiston perusteella halutaan yleistää uusiin havaintoihin. On oleellista, voidaanko luottaa, että uudet havainnot tulevat samasta jakaumasta. Jos voidaan, niin aineistosta opittujen säännönmukaisuuksien perusteella voidaan ennustaa. Jollei, niin ennustaminen on paljon hankalampaa, ja siksi useimmiten pitäydytään aiemman datan piirissä.

Se ei kuitenkaan tee päättelystä datan sisällä mitenkään triviaalia ja yksinkertaista. Esimerkiksi tiedonlouhinnassa haetaan datasta kuvioita ja säännönmukaisuuksia.

– Silloin voidaan esittää kysymys, että onko jokin kuvio oikeasti olemassa vai näyttääkö se vain siltä, vaikka minkäänlaisia ennusteita ei edes yritettäisi tehdä. Tähän on olemassa tehokas työkalu ja se on todennäköisyysperustainen mallitus.

Tutkimuksessa yhdistyvät tietojenkäsittelytiede, data-analyysi ja tilastotiede. Mikään niistä ei ole toista tärkeämpi tai keskeisempi, vaan ne kaikki yhdessä muodostavat kokonaisuuden, jonka varaan laskennallinen päättely rakentuu.

– Yksikön ytimessä on koneoppiminen. Se on käytännössä edistynyttä tilastotiedettä, jossa on otettava huomioon myös tietojenkäsittelylliset rajoitusehdot. Toisaalta se on tietojenkäsittelyä, jossa pyritään mallituksen kautta perusteltuihin algoritmeihin. Ja kun datasta puhutaan, niin periaatteessa kaikki on data-analyysiä. Koneoppimisessa juuri se on oikeastaan mullistavinta: kun nämä kolme asiaa on tuotu yhteen, on saatu aikaan yhdistelmä, jolla pystytään ratkaisemaan ongelmia, joita on aina haluttu ratkaista.

Vaikka virallisesti on kyse ”Suomalaisen laskennallisen päättelyn huippuyksiköstä”, Kaski käyttää laskennallisen päättelyn sijasta mieluummin nimitystä koneoppiminen.

– Koneoppimista voidaan käyttää hämmästyttävän monessa paikassa. Syynä on se, että yhä useampi ala on nykyisin datalähtöinen: dataa kerätään, se esitetään digitaalisessa muodossa ja siitä kootaan tietokantoja, joita voi käyttää erilaisiin tarkoituksiin. Riippumatta siitä, onko kyse humanistisesta tutkimuksesta, biologiasta, neurotieteistä tai ilmakehätutkimuksesta, on merkittävä kilpailuetu, että näitä tietokantoja osataan hyödyntää mahdollisimman monipuolisesti. Oikeastaan olisikin helpompi listata asiat ja alat, joilla koneoppimista ei nykypäivänä tarvittaisi tai voitaisi hyödyntää.

Huippuyksikössä on kuitenkin keskitytty laskennalliseen biologiaan ja lääketieteeseen.

Genomiaineistojen perusteella voidaan tehdä sairausdiagnooseja ja -prognooseja, neurotieteessä aivokuvantamismenetelmillä saadaan suunnattoman suuria tietokantoja, joita on pystyttävä analysoimaan.

– Humanistiselta puolelta voi ottaa esimerkiksi laskennallisen historian. Sillä voidaan tutkia vaikkapa kansansatujen tai uskonnollisten tekstien kehittymistä. Dataa on kertynyt aikojen saatossa siten, että munkit ovat kopioineet tekstejä. Aina kun on tehty virhe, seuraava versio on erilainen riippuen siitä, onko se kopioitu virheellisestä vai virheettömästä edeltäjästä. Näin muodostuu eräänlainen haaroittuva ”puu”, josta pystytään laskennallisesti päättelemään, mitkä versiot ovat varhaisimpia. Ja tästä päästäänkin yllättäen lähelle evoluutiomekanismeja. Samankaltaisilla algoritmeilla voidaan päätellä satujen eri versioiden esiintymisjärjestystä ja bakteerien evolutiivista kehittymistä. Jälkimmäinen on puolestaan tärkeää, jotta pystytään kehittämään tehokkaita antibiootteja.

Datan hyödyntäminen ja sen analyysimenetelmien kehittäminen saa aikaan myös eräänlaisen takaisinkytkentäilmiön: kun koneoppimista sovelletaan laajoihin tietokantoihin ja tähdätään tiettyihin tuloksiin, samalla saadaan vihiä siitä, millaista datan pitäisi olla ja miten sitä pitäisi koota, jotta sitä voitaisiin hyödyntää entistä tehokkaammin ja monipuolisemmin.

– Data-analyysia voi pitää tavallaan modernina mikroskooppina. Kaikki uudet mittausvälineet ovat aina muokanneet kulloistakin tieteenalaa, koska ne ovat tehneet mahdolliseksi uusien asioiden tarkastelun. Kun esimerkiksi mikroskooppi keksittiin, alettiin kokeita tehdä siten, että sillä pystyttiin todentamaan tehtyjä hypoteeseja. Sama pätee data-analyysiin: jos on jokin toimiva ja luotettava tapa analysoida dataa, sitä pyritään keräämään niin, että soveltamalla tätä menetelmää saadaan luotettavia tuloksia.

Yksi huippuyksikön tavoitteista on juuri datan ja sen analyysin jatkuva vuorovaikutus. Pyrkimyksenä on kehittää uusia menetelmiä, joiden avulla pystytään esittämään uusia ja tärkeitä kysymyksiä, joita ei aiemmin ole voitu kysyä, koska ei ole osattu joko mitata tai analysoida mittaustuloksia.

– Käynnissä on oikeastaan kaksi rinnakkaista sykliä. Ensinnäkin data-analyysimenetelmien kehittämisessä on teoria ja sitten ”havainto”, eli kun joku käyttää tiettyä menetelmää datan analysoimiseen, saadaan tietoa data-analyysimenetelmien edelleenkehittämistä varten. Toisessa syklissä ovat kunkin tieteenalan omat teoriat ja havainnot. Parhaimmillaan nämä syklit kulkevat tasatahtia, mutta toisinaan toinen ottaa pidemmän loikan ja mullistaa samalla toisenkin.

Tietyllä tavalla Kasken johtaman huippuyksikön edeltäjä oli ”Adaptiivisen informatiikan tutkimuksen huippuyksikkö”, jota johtanut professori Erkki Oja on mukana myös nykyisen huippuyksikön tutkimustyössä. Kyse ei kuitenkaan ole saman tutkimuksen jatkamisesta uudella nimellä.

– Adaptiivisen informatiikan yksikössä oli kehitetty erittäin hyviä puhtaasti datalähtöisiä malleja, joihin ei tuoda juurikaan etukäteisoletuksia eikä tietoa systeemistä. Tämä tutkimus otettiin uuden huippuyksikön yhdeksi alkupisteeksi, mutta pääasiaksi otettiin kaksi uutta teemaa ja niiden tutkimista varten otettiin mukaan tarvittavat tutkimusryhmät. Toinen on entistä vaikeampien ongelmien ratkaisu entistä monimutkaisempien mallien avulla, mukaanlukien useamman toisiinsa liittyvän aineiston käyttö.

Toinen teema on skaalautuminen. Datamäärän jatkuva kasvu edellyttää päätelmien tekemistä yhä suurempien aineistojen pohjalta, mutta toisaalta myös entistä nopeampien päätelmien tekemistä siten, että systeemistä saadaan interaktiivinen.

– Kiinnostava teema on myös, että koko ajan kehitetään uudenlaisia instrumentteja ja esitetään uudenlaisia kysymyksiä, joiden pohjalta tehdään mittauksia. Silloin kyseessä onkin jo datajoukkojen joukko ja haasteena on saada selville, mitä yhteyksiä näillä joukoilla on.

Huippuyksikön tutkimuksessa yksi keskeinen tavoite on kehittää data-analyysimenetelmiä, joilla pystytään hallitsemaan isoja kokonaisuuksia, datajoukkojen kokoelmia. Yksi käytännön esimerkki on juuri uusien mittalaitteiden tuottama data, joka kertoo osittain samoista asioista kuin aikaisemmatkin aineistot, mutta osittain

myös uusista asioista.

– Vaihtoehtoina on kehittää täsmällinen malli siitä, miten mittausaineistot liittyvät fysikaalisesti toisiinsa – tällaista ylellisyyttä ei läheskään aina ole – tai kysyä data-analyyttisesti, mitä yhteistä aineistoilla on. Kun kootaan mittauspareja, voidaan kehittää datasta oppiva tekniikka, joka pystyy kertomaan, mitä yhteistä näillä pareilla on.

Yksikössä on kehitetty esimerkiksi malleihin perustuva aineistojen hakuperiaate. Lähtökohtana oli solun toimintaan liittyvä molekyylibiologinen mittausaineisto. Jos haluaa selvittää, onko joku tehnyt samanlaisia mittauksia tai tutkinut samaa kysymystä, toistaiseksi ainoa keino löytää vastaus on kuvata tutkimuskysymys ja toivoa, että jossakin aiemmassa tutkimuksessa on käytetty kuvaukseen täsmälleen samoja sanoja.

– Me muotoilimme tähän tarkoitukseen mallien hakukoneen. Kun uudesta joukosta on olemassa datalähtöinen malli, on mahdollista kysyä, onko malleissa jotain yhteistä. Koska mallinnuksessa pyrkimyksenä on tiivistää datasta olennaiset asiat, malleja vertaamalla voidaan nähdä, löytyykö näissä olennaisissa asioissa yhteyksiä. Näyttää siltä, että hakuperiaate toimii ja se tosiaan löytää kiinnostavia aineistoja.

Periaate on laajennettavissa ja yleistettävissä myös tutkimusmaailman ulkopuolelle. Huippuyksikössä on pohdittu, miten koneoppiminen voisi mahdollisimman hyvin auttaa käyttäjää tehtävissä, joissa käyttäjä haluaa ja tarvitsee apua. Yksi sellainen on nimenomaan tiedonhaku.

– Kehitimme järjestelmän, joka seuraa käyttäjän tekemisiä ja pyrkii siltä pohjalta ennustamaan, mitä käyttäjä on hakemassa. Koneen kannalta ongelma on se, ettei sillä ole juurikaan tietoa siitä, mitä käyttäjä tekee tai mikä häntä kiinnostaa: yksi hakusana ei monimutkaisissa tehtävissä vielä paljon kerro. Nykyiset hakukoneet eivät osaa auttaa, jos käyttäjä joutuu esimerkiksi hakiessaan samalla opiskelemaan sitä, mitä oikeastaan onkaan hakemassa.

Ratkaisua haetaan interaktiivisesta tavoitteiden mallinnuksesta. Siinä systeemi pyrkii ennustamaan käyttäjän tavoitteita ja kiinnostuksia, ja näyttää ne tavanomaisten hakukoneosumien lisäksi.

– Kehitimme hakukoneprototyypin SciNet, jonka ensimmäisessä versiossa on käyttöliittymä nimeltä IntentRadar. Se on eräänlainen tutka, joka näyttää käyttäjää mahdollisesti kiinnostavia asioita. Käyttäjä voi antaa nopeasti palautetta ennusteista siirtämällä onnistuneet keskemmälle tutkaa ja ”hudit” kauemmas. Taustalla toimiva koneoppimisalgoritmi pyrki sitten palautteen perusteella päättelemään, mikä käyttäjää oikeasti kiinnostaa. Samalla se kuitenkin tarjoaa myös vähemmän ilmeisiä asioita, sillä käyttäjä jäisi eräänlaiseen ”kuplaan”, jos kone tarjoaisi vain hyviltä vaikuttavia osumia.

Tiedon suuri määrä on yksi keskeinen muutoksia aiheuttava tekijä. Aiemmin superlaskentaa käytettiin usein tehtävissä, joihin ei liittynyt suuria aineistoja, esimerkiksi simulaatiomallien laskennassa. Toisaalta kaikki data-analyysi ei ole vaatinut suunnatonta laskentatehoa.

– Seuraava suuri mullistus on tulossa, kun suuret datamäärät ja monimutkaiset mallit yhdistetään. Tällä hetkellä pullonkaulan voi ajatella olevan sekä datan että laskentatehon puolella. Kun jommassa kummassa tehdään uusi läpimurto, se auttaa myös toista pääsemään eteenpäin.

Nykyiset mittaukset tuottavat suoraan digitaalista aineistoa, mutta iso osa etenkin historiaan liittyvästä materiaalista on analogisessa muodossa arkistojen kätköissä. Miten laskennallinen päättely puree siihen?

– Suurin ongelma on todennäköisesti aineistojen ymmärtämisessä. Koska tiedon määrä kasvaa kaiken aikaa eksponentiaalisesti, kaukana menneisyydessä sitä on koko ajan suhteessa vähemmän eli datamassa ei ole kovin suuri. Aineistojen digitointi ei siis sinänsä ole mikään ylivoimainen urakka, mutta mitä vähemmän aineistoa on, sitä enemmän ammattitaitoa ja asiantuntemusta tarvitaan sen ymmärtämiseksi ja saattamiseksi käyttökelpoiseen muotoon.

Ks. myös kirjoitus ”Digitaaliset ihmistieteet tutkimuskartalle” (s. 29–32).

Kirjoittaja on tiedetoimittaja ja tietokirjailija.