Tietojenkäsittelytiede **34** (Huhtikuu 2012) sivut 6–16
Toimittaja: Jorma Tarhio

Globaali optimonti haastaa lokaalit menetelmät: luotettavampia riippuvuuksia tehosta tinkimättä

Wilhelmiina Hämäläinen
Itä-Suomen yliopisto
Biologian laitos
whamalai@cs.uef.fi

Työterveystarkastuksessa selvisi, että kolesterolisi on koholla. Myös verenpaineesi on melko korkea, mutta arvelet sen johtuvan vain stressistä. Lääkäri kuitenkin kehotti ottamaan asian vakavasti, sillä sekä korkea kolesteroli että verenpaine lisäävät ateroskleroosin (valtimotaudin) riskiä, mikä altistaa mm. sydän- ja aivoinfarktille. Lääkäri olisi määrännyt sinulle kolesterolilääkkeitä, mutta et haluaisi aloittaa lääkitystä ilman tarkkaa harkintaa. Niinpä testautat riskigeenisi. Selviää, että sinulla on ABCA1-geenin alleeli R219K sekä ApoE-geenin pahamaineinen alleeli ε4. Mutta mitä se merkitsee? Mitä sinun pitäisi tehdä?

Konsultoit kirjallisuutta ja lääketieteeseen perehtyneitä tuttavia, mutta kaikkialta saat eri vastauksia. Erään tutkimuksen mukaan ApoE-ε4 altistaisi nimenomaan aivoinfarktille, mutta toisen tutkimuksen mukaan niiden välillä ei olisi mitään riippuvuutta. Kolmannen tutkimuksen mukaan ApoE-ε4 ja vähäinenkin alkoholin käyttö olisivat erityisen vaarallinen yhdistelmä aivoinfarktin kannalta, mutta geenitutkijaystäväsi puoltaa kohtuullista alkoholinkäyttöä sydämen suojelemiseksi. Hän kehottaa sinua tutkituttamaan myös salaperäisen lipoproteiini(a):n (Lp(a)), jotta osaisi arvioida dementiariskisi. Itse harkitset tupakoinnin aloittamista, sillä kohtuullinen tupakointi näyttäisi suojelevan ApoE-ε4:n kantajia Alzheimerin taudilta. Toisaalta, jos et polta, sinun pitäisi ehkä alkaa rajoittaa kahvin juontia – tupakoijillahan se ei vaikuta riskeihin.

ABCA1-R219K:sta saat vielä ristiriitaisempaa tietoa. Tutkimuksesta riippuen se saattaa korreloida joko korkean tai matalan HDL-kolesterolipitoisuuden kanssa, estää tai ehkäistä sydänsairauksia, aivoinfarktia tai Alzheimerin tautia. Hoitosuositukset ovat aivan yhtä sekavia. Neuvonantajasi suosittelevat statiineja, niasiinia, C-vitamiinia, punaviiniä, suklaata, kasvissyöntiä, kalaöljyä – tai vain odottamaan, kunnes alat kaljuuntua. Lopulta suivaannut ja tivaat tutkijaystävältäsi, eikö nykyinen lääketiede tosiaan osaa antaa yksiselitteisiä vastauksia. Ystäväsi vastaa, ettei vika suinkaan ole lääketieteen tutkijoissa, vaan tietojenkäsittelijöissä, jotka eivät ole kehittäneet työkaluja heidän tarpeisiinsa. Ateroskleroosin, kuten monien muidenkin tautien, syntymekanismi on hyvin monimutkainen, eikä lääketieteen tutkijoilla yksinkertaisesti ole laskennallisia työkaluja analysoida kymmenien tuhansien geenien, fenologisten muuttujien ja ympäristötekijöiden yhteisvaikutuksia sairauksien synnyssä ja ehkäisyssä.

OHJE KIRJAPAINOLLE: B5-arkin vasen- ja yläreuna kohdistetaan A4-arkin vasempaan ja yläreunaan. Näin pitäisi marginaaliksi tulla taitteen puolella noin 33 mm ja muualla noin 22 mm.

Taulukko 1: Esimerkkejä sairauksiin liittyvistä tilastollisista riippuvuuksista. Lisää riippuvuussääntöjä sekä viitteet tutkimuksiin löytyvät sivulta `http://www.cs.uef.fi/~whamalai/esim.html`.

korkea LDL → ateroskleroosi
korkea HDL → ei ateroskleroosi
korkea Lp(a) → ateroskleroosi
tupakoi → ateroskleroosi
paljon tupakkaa → alzheimer
kahvi, ei tupakoi → ateroskleroosi
kohtuudella alkoholia → ei sydänsairaus
miestyypin kaljuus -> sydänsairaus
ApoE-e4 -> aivoinfarkti
ApoE-e4, kohtuudella alkoholia → alzheimer
ApoE-e4, kohtuudella tupakkaa → ei alzheimer
ApoE-e4, korkea Lp(a) → alzheimer
ABCA1-R219K → ei alzheimer
ABCA1-R219K, nainen → alzheimer

1 Tilastolliset riippuvuussäännöt

1.1 Mitä tilastollinen riippuvuus kertoo?

On tuskin liioittelua väittää, että empiiristen tieteiden perustehtävä on tilastollinen riippuvuusanalyysi. Juuri tilastollisten riippuvuuksien pohjalta voidaan muodostaa hypoteeseja kausaalisista syyseuraussuhteista ja oppia ymmärtämään luonnonlakeja ja säännönmukaisuuksia.

Tilastollinen riippuvuus ei tietenkään vielä itsessään merkitse kausaalisuhdetta, mutta se voi opastaa kausaalisuhteen jäljille. Esimerkiksi miestyypin kaljuuden ja sydänsairauksien välinen riippuvuus tuskin merkitsee, että kaljuus mitenkään rasittaisi sydäntä tai verisuonia. Sen sijaan tutkijat arvelevat, miestyypin kaljuus ja ateroskleroosi johtuisivat samasta tekijästä, kuten kohonneesta androgeenipitoisuudesta [13].

Kausaalisuhteiden löytäminen ja todistaminen on kuitenkin vaikeaa ja aikaavievää, eikä esimerkiksi ateroskleroosin tarkkaa syntymekanismia tunneta. Siitä huolimatta tunnetut tilastolliset riippuvuudet ovat arvokkaita päättelyvälineitä. Esimerkiksi korkean LDL-kolesterolin ja ateroskleroosin välinen riippuvuus on niin vahva, että sen pohjalta kannattaa tehdä hoitosuosituksia. Geneettisten tekijöiden ja sairauksien väliset riippuvuudet ovat monimutkaisempia, mutta tiettyjen alleelien (geenivarianttien) perusteella voidaan tunnistaa riskiryhmiin kuuluvia henkilöitä.

1.2 Riippuvuuksien esitys sääntöinä

Binääristen muuttujien väliset riippuvuudet voidaan esittää havainnollisesti *riippuvuussääntöinä*. Taulukossa 1 on annettu esimerkkejä sydän- ja verisuonitauteihin sekä Alzheimerin tautiin liittyvistä riippuvuussäännöistä. Esimerkiksi sääntö *tupakka → ateroskleroosi* kertoo, että tupakoinnin ja ateroskleroosin välillä on positiivinen riippuvuus (tupakointi lisää ateroskleroosin riskiä) ja vastaavasti tupa-

koimattomuuden ja ateroskleroosin välillä on negatiivinen riippuvuus (tupakoimattomuus pienentää riskiä).

Moniarvoiset kategoriset muuttujat voidaan aina helposti binärisoida. Esimerkiksi ApoE-geeniparin kaikille mahdollisille alleelikombinaatioille voidaan luoda omat binäärimuuttujat: ε2ε2, ε2ε3, ε2ε4, ε3ε3, ε3ε4, ε4ε4. Käytännössä alleelilla ε4 on niin voimakas vaikutus, että arvot ε3ε4 ja ε4ε4 on tapana yhdistää yhdeksi binäärimuuttujaksi ApoE-ε4. Suomalaisten kannalta tämä muuttuja on erityisen kiinnostava, sillä noin kolmanneksella meistä on vastaava alleelikombinaatio, minkä on arveltu osaltaan selittävän suurta sydän- ja verisuonitautien esiintyvyyttä [10].

Myös numeerisia muuttujia voi tarkastella riippuvuussäännöissä, kunhan ne ensin diskretoidaan ja muunnetaan sopiviksi binäärimuuttujiksi. Esimerkiksi tupakan kulutus (askia päivässä) voidaan jakaa kolmeen osaväliin: vähän (alle aski), kohtuudella (1–2 askia) tai paljon (yli 2 askia) tupakkaa päivässä. Tällainen muunnos hävittää aina informaatiota ja valittu osavälijako vaikuttaa löydettäviin riippuvuuksiin. Toisaalta muunnos poistaa myös haitallista satunnaisvaihtelua, mikä voi auttaa allaolevan riippuvuustrendin paljastamisessa.

Riippuvuussääntöjä voidaan yksinkertaistaa lisää vaatimalla, ettei säännön ehto-osassa saa esiintyä negaatioita. Tämä ei rajoita mahdollisten riippuvuuksien avaruutta, mikäli samalla luomme uudet ekstramuuttujat kaikille negaatioille. Sivutuotteena onnistumme esittämään myös arvojen disjunktioita. Esimerkiksi muuttuja ”ei paljon tupakkaa” vastaa disjunktiota ”vähän tupakkaa” tai ”kohtuudella tupakkaa”.

Toisaalta on monia sovellusalueita, joissa tietyn arvon läsnäolo on paljon kiinnostavampaa kuin sen poissaolo, eikä negaatiomuuttujia tarvitse luoda. Esimerkiksi jos jokin alleeli esiintyy vain parilla prosentilla ihmisistä, ei sen puuttuminen (oletusarvo) kerro paljoakaan. Tällainen strategia saattaa olla laskennallisestikin välttämätöntä, koska usein esiintyvien muuttujien välillä on aina paljon sattumasta johtuvia *näennäisriippuvuuksia* (spurious dependency) ja muiden riippuvuuksien sivutuotteena syntyneitä *redundatteja riippuvuuksia* (redundant dependency). Tällaisten riippuvuuksien hakuun on turha käyttää laskentaresursseja, koska ne eivät kerro mitään uutta informaatiota aidosti pätevistä riippuvuuksista.

1.3 Aidon riippuvuussäännön kriteerit

Kirjoittajan väitöskirjatutkimuksessa *aidolle riippuvuussäännölle* (genuine dependency rule) esitettiin kolme vaatimusta, jotka käyvät yksiin sovellusalueilla esitettyjen vaatimusten kanssa (esim. [12]). Sen lisäksi, että sääntö ilmaisee tilastollista riippuvuutta, sen tulisi olla tilastollisesti merkitsevä ja ei-redundantti.

Määritelmä 1 (Aito riippuvuussääntö)**.** *Olkoon R joukko binäärimuuttujia, $X \subseteq R$, $A \in R \setminus X$ ja $a \in \{0,1\}$. Sääntö $X \to A = a$ (eli joko $X \to A$ tai $X \to \neg A$) on aito riippuvuussääntö, mikäli seuraavat ehdot pätevät:*

1. *$X \to A = a$ ilmaisee tilastollista riippuvuutta eli $P(XA = a) \neq P(X)P(A = a)$. Koska A on binäärinen, riittää esittää vain positiiviset riippuvuudet ($P(XA = a) > P(X)P(A = a)$), sillä säännön komplementti $X \to A \neq a$ ilmaisee vastaavan negatiivisen riippuvuuden.*

2. *Riippuvuus on tilastollisesti merkitsevä. Tämän arvioimiseksi lasketaan todennäköisyys (”p-arvo”), että havaitun vahvuinen riippuvuus esiintyisi da-*

tassa sattumalta, mikäli X ja A olisivat toisistaan riippumattomia. Mitä pienempi todennäköisyys on, sitä uskottavampaa on, että kyseessä on aito riippuvuus, joka pätee myös tulevassa datassa. (Tarkan p-arvon sijasta voidaan toki käyttää muitakin tilastotieteen, informaatioteorian, koneoppimisen ja tiedonlouhinnan riippuvuuksille kehittämiä merkitsevyys- ja hyvyysmittoja.)

3. *Riippuvuus ei ole redundantti eli jonkun yleisemmän riippuvuuden sivutuotetta ilman lisäinformaatiota. Toisin sanoen $X \to A = a$ on parempi (käytetyllä hyvyysmitalla) kuin mikään sen yleistys $Y \to A = a$, $Y \subsetneq X$.*

Tilastollisen merkitsevyyden arvioinnissa käytettävä mitta riippuu valitusta tulkintamallista. Sopivaa mallia valittaessa tulee ennen kaikkea päättää, tutkitaanko binäärimuuttujien X ja A välistä riippuvuutta (muuttujaperustainen tulkinta) vai ainoastaan tietyn arvokombinaation $X = x, A = a, x, a \in \{0,1\}$, välistä riippuvuutta (arvoperustainen tulkinta). Kummassakin tapauksessa on käytettävissä useita vaihtoehtoisia malleja riippuen siitä, mitä oletuksia tehdään. Näitä on käsitelty perusteellisesti väitöskirjan [7] luvussa 2.

Tunnetuimpia tilastollisen merkitsevyyden mittoja ovat χ^2-mitta

$$\chi^2(X \to A) = \frac{n(P(X,A) - P(X)P(A))^2}{P(X)P(\neg X)P(A)P(\neg A)}$$

sekä Fisherin eksaktin testin p-arvo

$$p_F(X \to A) = \sum_{i=0}^{J} \frac{\binom{fr(X)}{fr(XA)+i}\binom{fr(\neg X)}{fr(\neg X \neg A)+i}}{\binom{n}{fr(A)}},$$

missä n on datan koko (rivien lukumäärä), fr absoluuttinen frekvenssi ja $J = \min\{fr(X\neg A), fr(\neg XA)\}$.

χ^2-mittaa on käytetty melko yleisesti optimaalisten luokittelusääntöjen etsinnässä, mutta Fisherin p-arvoa ei ole aiemmin käytetty missään (oikeellisesti toimivassa) hakualgoritmissa. Tämä on vahinko, sillä kirjoittajan tutkimusten mukaan Fisherin p-arvo tuottaa selvästi luotettavampia tuloksia, ts. löydetyt riippuvuudet pätevät paremmin tutkimusjoukon ulkopuolisessa datassa. Tämä ei sinänsä ole yllättävää, sillä χ^2-mitta on hyvin herkkä datan jakaumalle [24, 2] ja Fisherin p-arvoa onkin suositeltu käytettäväksi aina, kun se on laskennallisesti mahdollista.

Analyyttisten mittojen lisäksi tilastollista merkitsevyyttä voidaan arvioida myös empiirisesti *satunnaistamistesteillä* (randomization tests, ks. esim. [6]). Ideana on generoida satunnaisia datajoukkoja, joissa pätevät jotkin alkuperäisen joukon piirteet, ja tutkia, kuinka suuressa osassa satunnaisjoukkoja esiintyy vähintään yhtä vahva riippuvuus kuin alkuperäisessä datajoukossa. Tämä menetelmä ei kuitenkaan sovellu riippuvuussääntöjen hakuun kuten analyyttiset mitat, vaan ainoastaan löydettyjen riippuvuuksien merkitsevyyden arviointiin. Ainut poikkeus on eksakti permutaatiotesti, jossa generoidaan kaikki mahdolliset datajoukot, joissa $fr(X)$ ja $fr(A = a)$ pätevät. Tällöin empiirinen p-arvo voidaan esittää analyyttisessä muodossa, joka on sama kuin Fisherin p_F.

Tässä yhteydessä on syytä huomauttaa, ettei löydettyjen riippuvuussääntöjen p-arvoista voi vielä tehdä päätelmiä siitä, ovatko riippuvuudet tilastollisesti merkitseviä klassisessa (Neyman-Pearsonilaisessa) mielessä. Klassisen hypoteesin testauksen ideanahan on kiinnittää jokin absoluuttinen raja-arvo α (tyypillisesti $\alpha = 0.05$ tai $\alpha = 0.01$) ja pitää löytöä tilastollisesti merkitsevänä, mikäli sen p-arvo ei ole α:aa suurempi. Taustaajatuksena on, että α kertoo todennäköi-

syyden, jolla riippumattomuutta ilmaiseva näennäissääntö voisi läpäistä virheellisesti testin. Riippuvuussääntöjen systemaattisessa haussa testataan kuitenkin niin monia potentiaalisia riippuvuuksia (ns. multiple testing problem), ettei α:n tulkinta enää päde. Tilastotieteilijät eivät ole saavuttaneet yksimielisyyttä siitä, miten ongelma pitäisi ratkaista tai ovatko ennaltakiinnitetyt α-arvot ylipäänsä järkeviä. Niinpä on turvallisinta käyttää *p*-arvoja vain hyvyysmittoina: niiden avulla voi löytää kaikkein merkitsevimmät riippuvuussäännöt, vaikka merkitsevyystasosta ei voikaan sanoa mitään.

Redundanssin tarkistus on tärkeää riippuvuuksien oikean tulkinnan kannalta, sillä redundantissa säännössä on ylimääräisiä muuttujia, jotka eivät todellisuudessa vaikuta riippuvuuteen. Mikäli ylimääräiset muuttujat heikentävät säännön konfidenssia, ts.
$P(A = a|X) < P(A = a|Y)$ jollain $Y \subsetneq X$,
on sääntö $X \to A = a$ varmasti redundantti. On kuitenkin mahdollista, että ylimääräiset muuttujat vahvistavat säännön konfidenssia sattumalta. Seuraava esimerkki havainnollistaa tätä.

Esimerkki. 100:n suomalaisen tietojenkäsittelytieteilijän joukosta 50 kärsi korkeasta verenpaineesta. 60 kärsi stressistä ja näistä 48:lla oli korkea verenpaine. Säännön *stressi* $\to$ *verenpaine* konfidenssi oli siis 80%. Riippuvuuden merkitsevyys Fisherin eksaktilla testillä oli $p_F = e^{-32.1}$ (eli $\ln(p_F) = -32$) eli sääntö oli hyvin merkitsevä. Datajoukosta löytyi kuitenkin vieläkin vahvempi riippuvuus: niiden 56:n stressistä kärsivän joukossa, jotka söivät säännöllisesti suklaata, peräti 46 kärsi korkeasta verenpaineesta. Säännön *stressi suklaa* $\to$ *verenpaine* konfidenssi oli siis 82%. Sääntö oli kuitenkin vähemmän merkitsevä, $p_F = e^{-30.6}$. Tämän perusteella pääteltiin, että se oli redundantti, eikä suklaa lisännyt merkitsevästi stressin ja verenpaineen välistä riippuvuutta. Mikäli sääntöjen merkitsevyyksiä ei olisi vertailtu, olisi esimerkiksi työterveysasema voinut olettaa, että verenpaineen tärkein riskiryhmä olivat suklaata syövät stressaantuneet työntekijät. Tämän perusteella olisi saatettu esimerkiksi sulkea laitoksen suklaa-automaatteja ja keskittää voimavaroja suklaan vastaiseen valistukseen.

2 Riippuvuussääntöjen haku

2.1 Laskennallinen ongelma

Riippuvuussääntöjen hakuongelma esiintyy kahdessa muodossa. *Optimointiongelmassa* tehtävänä on etsiä *K* parasta riippuvuussääntöä kaikkien mahdollisten riippuvuussääntöjen joukosta, kun taas *luettelointiongelmassa* (enumeration problem) tehtävänä on etsiä kaikki riittävän hyvät riippuvuussäännöt, annettuna jokin hyvyysmitan ala- tai yläraja (*p*-arvolle annetaan yläraja ja χ^2-mitalle alaraja). Lisäksi edellytetään, että löydetyt riippuvuussäännöt ovat ei-redundantteja.

Ongelma on laskennallisesti erittäin vaativa, sillä jo yksinkertaisempi ongelma – parhaan luokittelusäännön $X \to C$ etsintä kiinnitetyllä luokkamuuttujalla C – on *NP*-kova tavanomaisilla hyvyysmitoilla [15]. Tämä ei ole sinänsä yllättävää, kun mietimme hakuavaruuden kokoa. Riippuvuussääntö voidaan muodostaa mistä tahansa vähintään kahden binäärimuuttujan joukosta X, ja mikä tahansa X:n muuttuja voi olla säännön seurausosana. Mikäli alkuperäinen muuttujajoukko R koostui k:sta muuttujasta, on mahdollisia riippuvuussääntöjä $\sum_{i=2}^{k} i \binom{k}{i} = k2^{k-1} - k$ kappaletta. Selvästikin hakuavaruuden koko kasvaa eksponentiaalisesti muuttujien lukumäärän k funktiona. Esi-

merkiksi jos $k = 40$, mahdollisia riippuvuussääntöjä on jo yli 20 biljoonaa. Analysoitavissa datajoukoissa binäärimuuttujia saattaa kuitenkin olla tuhansia tai jopa kymmeniä tuhansia.

Riippuvuussääntöjen hakua mutkistaa se, että tilastollinen riippuvuus *ei ole antimonotoninen ominaisuus*. Tämä merkitsee sitä, että $X \to A = a$ voi ilmaista tilastollista riippuvuutta, vaikka kaikki sen yleistykset $Y \to A = a, Y \subsetneq X$, ilmaisisivat vain riippumattomuutta. Yhtä lailla $X \to A = a$ voi ilmaista positiivista riippuvuutta, vaikka kaikki sen yleistykset ilmaisisivat negatiivista riippuvuutta, tai päinvastoin. Sama ongelma koskee tilastollisia merkitsevyysmittoja: $X \to A = a$ voi olla riittävän merkitsevä, vaikka yksikään sen yleistyksistä ei olisi.

Juuri antimonotonisuuden puute on tärkeimpiä syitä geenitutkimuksessa saataviin ristiriitaisiin tietoihin geenien ja sairauksien riippuvuuksista. Koska skaalautuvia työkaluja ei ole ollut, on käytäntönä ollut esivalita lupaavimman tuntuisia alleeleja sen perusteella, miten hyvin ne yksinään korreloivat sairauden tai tutkittavan fenologisen muuttujan kanssa. Kun kymmenistä tuhansista potentiaalisesti relevanteista alleeleista valitaan korkeintaan muutamia satoja, on erittäin suuri riski, että oikeasti tarpeellisia muuttujia jää datajoukon ulkopuolelle. Nykyisin arvellaankin, että esimerkiksi syöpien, ateroskleroosin ja diabeteksen syyt ovat hyvin monimutkaisia, eivätkä löydetyt yksinkertaiset riippuvuudet selitä niistä murto-osaakaan. [5, 14] Siksi olisi ensiarvoisen tärkeää päästä eroon ahneesta muuttujien valinnasta, joka voi tuottaa hyvinkin harhaanjohtavia tuloksia.

2.2 Aiempia ratkaisuja sukulaisongelmiin

Kirjoittajan väitöskirjatutkimuksen haasteena oli kehittää tehokkaita algoritmeja sekä optimointi- että luettelointiongelmaan aitojen riippuvuussääntöjen löytämiseksi. Keskeisenä vaatimuksena oli, ettei hakua saisi rajoittaa millään epäoptimaalisilla heuristiikoilla, vaan löydettyjen sääntöjen tulisi todella olla parhaita, mitä kyseisestä datajoukosta voi löytää. Yllättävää kyllä, aiemmasta tutkimuksesta ei löytynyt yhtään algoritmia tähän ongelmaan. Siksi seuraavassa on kuvattu tehokkaimpia samantapaisiin ongelmiin kehitettyjä algoritmeja. Kaikki kuvatut algoritmit kykenevät etsimään tilastollisia riippuvuuksia, mutta riippuvuudet eivät välttämättä ole ei-redundantteja tai merkitsevimpiä, mitä datasta löytyy. Mikään kuvatuista menetelmistä ei myöskään takaa globaalia optimia.

Suurin osa aiemmasta lähialueiden tutkimuksesta on keskittynyt *frekventtien joukkojen* ja niistä johdettavien *assosiaatiosääntöjen* [1] etsintään. Perusideana on etsiä ensin kaikki riittävän frekventit joukot $Z \subseteq R$ jollain minimifrekvenssillä. Kustakin joukosta Z voidaan sitten muodostaa sääntöjä $Z \setminus \{A\} \to A$ jollain kriteerillä. Perinteisesti sääntöjen kriteerinä on käytetty minimikonfidenssia, mutta se ei takaa tilastollista riippuvuutta. On jopa mahdollista, että sääntö ilmaisee negatiivista riippuvuutta, vaikka käyttäjä olettaa löytävänsä positiiviset riippuvuudet. Sääntöjä voidaan kuitenkin karsia lisää vaatimalla riittävän vahvaa tilastollista riippuvuutta (mittoina tyypillisesti *lift*, $\gamma(X \to A) = \frac{P(XA)}{P(X)P(A)}$, tai *leverage*, $\delta(X \to A) = P(XA) - P(X)P(A)$) tai riittävää tilastollista merkitsevyyttä. Lisäksi löydetyistä jopa miljoonista säännöistä tulisi karsia redundantit säännöt, mikä voi olla hyvinkin aikaavievää.

Lähestymistavan pahin ongelma on huono skaalautuvuus. Mikäli data on yhtään tavanomaista ostoskoridataa tiheämpää (ts. paljon 1-arvoisten muuttujien kombinaatioita), täytyy minimifrekvenssin olla varsin suuri. Pahimmassa tapauksessa vaadittava minimifrekvenssi voi olla luokkaa 60–80%, mikä merkitsee, että kaikki löydetyt riippuvuudet ovat hyvin yleisiä (esiintyvät vähintään 60–80%:lla riveistä) ja siten myös heikkoja. Usein käykin niin, ettei kaikkein merkitsevimpiä riippuvuuksia löydetä lainkaan [21].

Huomattavasti pienemmät minimifrekvenssit ovat mahdollisia, mikäli etsitään vain jonkinlainen tiivistetty esitys frekventeistä joukoista, esimerkiksi kaikki *suljetut joukot* (closed sets). Riippuvuussääntöjen etsinnässä tällaisista tiivistetyistä esityksistä ei kuitenkaan ole mitään hyötyä. Esimerkiksi suljetut joukot sisältävät vain kaikkein spesifeimmät tietyn frekvenssin omaavat joukot (ts. jos X on suljettu, ovat kaikki $Y \subsetneq X$, joille $P(X) = P(Y)$, ei-suljettuja) ja niistä johdetut säännöt sisältävät useimmiten redundantteja muuttujia (mikäli ne eivät sisältäisi yhtään redundantteja muuttujia, olisivat kaikki frekventit joukot suljettuja, eikä esitys tiivistäisi mitään). Jopa kymmenien redundanttien muuttujien tunnistus ja eliminointi jälkikäteen on kuitenkin aikaavievää. Ongelmasta on kerrottu lisää väitöskirjan [7] liitteessä B.2.

Edellistä huomattavasti tehokkaampi lähestymistapa on etsiä suoraan halutunlaisia assosiaatio- tai riippuvuussääntöjä, vaikka haku perustuisikin minimifrekvenssiin. Tämän lähestymistavan huomattavin edustaja on Webb, jonka MagnumOpus-ohjelmistolla voi etsiä riippuvuussääntöjä mm. lift- ja leverage-mitoilla [20, 22]. Sääntöjen tilastollista merkitsevyyttä ei tutkita, mutta sen sijaan on mahdollista varmistaa, että säännön parannus suhteessa sen välittömiin yleistyksiin ($X \to A$ suhteessa kaikkiin $X \setminus \{A_i\} \to A, A_i \in X$) on tilastollisesti merkitsevä. Algoritmi on niin tehokas, että haku (ainakin leverage-mitalla) onnistuu useimmissa datajoukoissa ilman mitään minimifrekvenssiä, kunhan muistia vain on riittävästi.

Luokittelusäännöille $X \to C$ (missä C on kiinnitetty luokkamuuttuja) on kehitetty useitakin algoritmeja, jotka käyttävät joko riippuvuuden vahvuutta tai tilastollista merkitsevyyttä hakukriteerinä. Yleensä algoritmit edellyttävät lisäksi muita rajoituksia kuten minimifrekvenssiä tai maksimaalista monimutkaisuutta (muuttujien lukumäärää X:ssä). Tavallisimmin hyvyysmittana on χ^2-mitta, kuten Morishitan ja Sesen [15] sekä Nijssenin et al. [17, 16] algoritmeissa. Lin algoritmi [11] poikkeaa sikäli muista, että se etsii vain parhaat ei-redundantit luokittelusäännöt erilaisilla tilastollisen vahvuuden mitoilla (lift, leverage). Minimifrekvenssiä ei tarvita, mutta sen sijaan käyttäjän pitää antaa raja-arvo $P(X|A)$:lle.

Lähes poikkeuksetta aiemmat hakualgoritmit etsivät vain positiivisia riippuvuussääntöjä. Tämä johtuu siitä, että negatiivisten riippuvuuksien etsintä on huomattavasti raskaampaa ja vaikeampaa, mikäli haku perustuu frekvenssipohjaiseen karsintaan. Poikkeuksia ovat Thiruvadyn ja Webbin [19] sekä Antonien ja Zaïanen [3] algoritmit, joista ensiksi mainittu käyttää leveragea ja jälkimmäinen Pearsonin korrelaatiokerrointa hyvyysmittana. Kumpikin algoritmi vaatii minimifrekvenssiä, eikä niiden skaalautuvuudesta ole tietoa.

Edellä mainittujen algoritmien lisäksi löytyy myös heuristisia algoritmeja, jotka perustuvat virheelliseen oletukseen, että tilastollinen riippuvuus (esim. [23]) tai tilastollinen merkitsevyys (esim. [9]) olisivat antimonotonisia ominaisuuksia. On-

gelmaa on analysoitu tarkemmin artikkelissa [8].

2.3 Kingfisher – tehostettu branch-and-bound

Kirjoittajan väitöskirjatutkimuksen haasteena oli kehittää tehokkaita hakualgoritmeja merkitsevimpien riippuvuussääntöjen hakuun erilaisilla merkitsevyys- ja hyvyysmitoilla. Sen sijaan tavoitteena ei ollut ratkaista poleemista ikuisuuskysymystä, miten riippuvuussäännön merkitsevyys tai hyvyys tulisi määritellä.

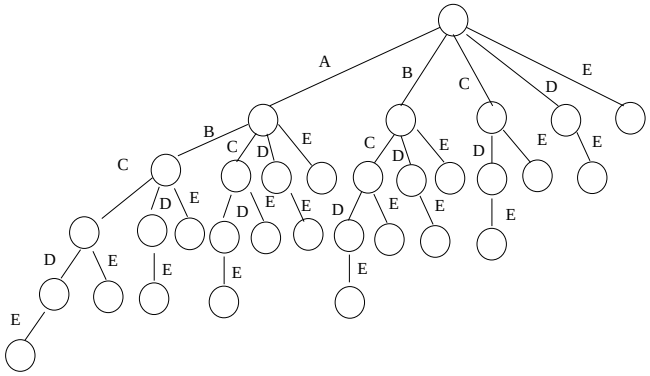
Tutkimuksen tuloksena syntyikin joukko erilaisia strategioita käyttäviä hakualgoritmeja erilaisille hyvyysmitoille. Väitöskirjaan niistä pääsi kuitenkin vain yksi, alkujaan Fisherin p-arvolle kehitetty *Kingfisher*. Se osoittautui ylivertaiseksi muihin nähden kahden tärkeän oivalluksen takia. Ensinnäkin selvisi, että kaikki *hyvinkäyttäytyvät mittafunktiot* saavat parhaat arvonsa (ylä- tai alarajansa) samoissa hakuavaruuden pisteissä kuin Fisherin p:kin. Hyvinkäyttäytyvät mittafunktiot on määritelty Piatetsky-Shapiron klassisten aksioomien [18] perusteella ja käytännössä kaikki intuitiivisesti järkevät riippuvuuksien hyvyysmitat kuuluvat tähän joukkoon. Tämän oivalluksen pohjalta oli helppo laatia geneerinen algoritmi, joka toimii millä tahansa hyvinkäyttäytyvällä mittafunktiolla.

Toinen oivallus oli algoritminen keksintö, joka tehosti todella merkittävästi allaolevaa branch-and-bound-strategiaa. Ihmeitä tekevän tehonsa vuoksi se sai nimen *Lapis Philosophorum -periaate*. Periaatteen yksityiskohtia on vaikea selostaa algoritmista irrallaan, joten kerron seuraavassa vain algoritmin ja karsintaperiaatteen perusidean. Lisää tietoa löytyy väitöskirjan [7] luvusta 4 sekä artikkelista [8], jossa on myös yksityiskohtainen simulaatio.

Kingfisher, kuten monet muutkin algoritmit, perustuu *luettelointipuun* (enumeration tree) läpikäyntiin. Kuvassa 1 on esitetty muuttujien $A,\ldots,E$ täydellinen luettelointipuu, joka esittää implisiittisesti kaikki muuttujajoukot $X \in \mathcal{P}(R)$. Solmua vastaava muuttujajoukko saadaan lukemalla juuri–solmu-polulla esiintyvät muuttujanimet. Solmussa X voidaan muodostaa säännöt $X \setminus \{A_i\} \rightarrow A_i = a_i$ ($A_i \in X$, $a_i \in \{0,1\}$).

Algoritmin perusideana on pitää jokaisessa solmussa kirjaa niistä $A_i = a_i$ ($A_i \in R$, $a_i \in \{0,1\}$), jotka ovat mahdollisia seurausosia kyseiselle joukolle X tai jollekin sen ylijoukolle XQ, $Q \subseteq R \setminus X$ (joukoille $X \setminus \{A_i\}$ tai $X \setminus \{A_i\}Q$, mikäli $A_i \in X$). Seurausosa on mahdollinen, mikäli vastaava sääntö voisi olla ei-redundantti ja riittävän merkitsevä, ja muuten se on mahdoton. Alussa kaikki seurausosat ovat mahdollisia, mutta haun edetessä niitä päivitetään mahdottomiksi. Kun solmun kaikki seurausosat ovat muuttuneet mahdottomiksi, ei vastaavalle joukolle luoda enää ylijoukkoja. Käytännössä haku on sitä tehokkaampaa, mitä enemmän ja mitä aikaisemmin seurausosia voidaan päivittää mahdottomiksi.

Kingfisherin edeltäjissä tieto seurausosan mahdollisuudesta siirtyi puussa vain ylhäältä alaspäin (kuitenkin myös haarojen välillä, sillä joukon X mahdolliset seurausosat saadaan leikkauksena joukkojen $Y \subsetneq X$, $|Y| = |X| - 1$, seurausosista). Lapis Philosophorumin taustaoivalluksena oli, että tietoa mahdollisista seurausosista voi ja kannattaa siirtää myös ylöspäin, edeltävälle tasolle. Kun luettelointipuun tasoa käydään vasemmalta oikealle, saadaan usein informaatiota, jonka perusteella voidaan poistaa mahdollisia seurausosia tulevien (oikeanpuoleisten) haarojen edelliseltä tasolta. Parhaimmillaan tämä johtaa siihen, että edellisen tason



Kuva 1: Täydellinen luettelointipuu muuttujista A, B, C, D, E.

solmuja saattaa hävitä kokonaan, ennen kuin ehditään niihin saakka. Tämä tuottaa merkittävää säästöä, sillä lehtitason koko on haun pullonkaula, niin aikavaativuuden kuin muistin kulutuksen kannalta.

Käytännössä Lapis Philosophorum-periaate muutti algoritmia radikaalisti tehokkaammaksi. Parhaimmillaan se pystyi karsimaan pullonkaulatason koon alle tuhannesosaan (Pumsb-joukko). Ilman Lapis Philosophorum -periaatetta ei negatiivisten riippuvuussääntöjen etsintä onnistunut monistakaan datajoukoista (Chess, Accidents, Pumsb) ilman minimifrekvenssiä edes Laskentakeskuksen supertietokoneilla (256 GB muistia). Lapis Philosophorum -periaatteen kanssa ohjelma kuitenkin pystyi tutkimaan kaikki testatut datajoukot ripeästi, kun mittana oli Fisherin p_F. χ^2-mitta sen sijaan hidasti hakua merkittävästi, eikä ohjelma pystynyt käsittelemään Pumsb-joukkoa järkevässä ajassa ilman pientä minimifrekvenssiä. Syynä oli se, että χ^2-mitalle ei voi asettaa tiukkoja ylärajoja – se saa maksimaalisen arvonsa aina, kun $P(XA = a) = P(X) = P(A = a)$. Siksi sen karsintateho on huono ja haku jatkuu usein hyvin syvälle. Valitettavasti myös sääntöjen laatu saattaa olla huono, sillä datasta löytyy usein monimutkaisia mutta hyvin harvinaisia sääntöjä, jotka saavat maksimaalisen χ^2-arvon. Silti ne saattavat olla sattuman tuotetta, eivätkä päde lainkaan testidatassa. Tässä mielessä ei ole lainkaan vahinko, ettei χ^2:lla voida aina etsiä optimaalisia sääntöjä ilman pientä minimifrekvenssiä.

Vertailun vuoksi mainittakoon, että samaisessa supertietokoneessa testattu assosiaatiosääntöohjelma [4] vaati niin suuria minimifrekvenssejä, ettei 100:aa merkitsevintä sääntöä olisi voinut löytää useimmista joukoista edes periaatteessa. Jopa klassinen ostoskoridatajoukko T40I10D100K osoittautui niin hankalaksi, ettei ohjelma pystynyt löytämään yhtään 100:sta merkitsevimmästä positiivisesta riippuvuussäännöstä.

Kingfisher-ohjelman eduksi on sanottava, ettei se suinkaan kaivannut Laskentakeskuksen suurta muistia, kun mittana oli p_F. Aivan kaikki testit pystyi ajamaan jopa vanhalla sylimikrolla (1 GB muistia) ilman mitään lisärajoituksia. Osittain tähän varmasti vaikutti vanha pedagoginen totuus: algoritmisuunnittelijan kannattaa toteuttaa algoritminsa vanhalla koneella, jotta niistä on pakko tehdä tehokkaita.

3 Tulevaisuuden tutkimushaasteet

Tutkimuksen tärkein huomio oli, ettei globaaleja optimointimenetelmiä kannata suotta väheksyä, vaikka ongelma olisikin kompleksinen. Eksponentiaalista aikavaativuudesta huolimatta voi aina keksiä tehokkaita ja skaalautuvia algoritmeja, jotka toimivat käytännössä vähintään yhtä hyvin kuin aiemmat heuristiset algorimit.

Riippuvuussääntöjen osalta tutkimushaasteet eivät suinkaan ole loppuneet. Ehkä kiinnostavin jatkotutkimushaaste on, miten käsitellä numeerisia muuttujia. Kuten alussa kerroin, vaikuttaa diskretointi löydettäviin riippuvuuksiin. Ideaalista olisi, jos optimaalisen diskretoinnin voisi löytää samalla kuin optimaaliset säännötkin. Toinen kiinnostava ongelma koskee tilastollisen riippuvuuden ja kausaalisuuden suhdetta. Vaikka tilastollisen riippuvuuden kausaalisuutta ei pystyisikään todistamaan, on olemassa tilanteita, joissa voi osoittaa riippuvuuden eikausaalisuuden kyseisen datan valossa.

Palatakseni vielä sairauksien syihin, toivoisin todellakin, että algoritmini voisi palvella lääketieteen ja muiden sovellusalueiden tutkijoita. Omien eksperimenttieni perusteella uskon nimittäin vakaasti, että algoritmini avulla olisi mahdollista löytää aivan uutta tietoa esimerkiksi sairauksien geneettisistä syistä. Sitä ennen Kingfisher tarvitsisi kuitenkin helppokäyttöisen käyttöliittymän tarvittavine pikku työkaluineen[1]. Komentoriviversion saa sivulta `http://www.cs.uef.fi/~whamalai/sourcecode.html`.

Kiitokset

Kiitän Tietojenkäsittelytieteen seuraa ja Tietotekniikan Tutkimussäätiötä osakseni tulleesta kunniasta.

[1] Mikäli olet kiinnostunut, ota yhteyttä kirjoittajaan.

Viitteet

1. R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM Press, 1993.
2. A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.
3. M.-L. Antonie and O.R. Zaïane. Mining positive and negative association rules: an approach for confined rules. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04)*, pages 27–38. Springer-Verlag, 2004.
4. C. Borgelt. Apriori v5.14 software, 2010. `http://www.borgelt.net/apriori.html`. Retrieved 7.6. 2010.
5. H.J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10:392–404, June 2009.
6. E.S. Edgington. *Randomization Tests*. Marcel Dekker, Inc., New York, third edition, 1995.
7. W. Hämäläinen. *Efficient Search for Statistically Significant Dependency Rules in Binary Data*. PhD thesis, Department of Computer Science, University of Helsinki, Finland, 2010. Series of Publications A, Report A-2010-2.
8. W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and Information Systems: An International Journal (KAIS)*, 2011. To appear. `http://www.cs.uef.fi/~whamalai/research.html`.
9. Y.S. Koh and R. Pears. Efficiently finding negative association rules without support threshold. In *Advances in Artificial Intelligence, Proceedings of the 20th Australian Joint Conference on Artificial Intelligence (AI 2007)*, volume 4830 of *Lectu-*

re Notes in Computer Science, pages 710–714. Springer, 2007.

10. T. Lehtimäki, T. Moilanen, T. Nikkari, T. Solakivi, K. Porkka, C. Ehnholm, T. Rönnemaa, H.K. Akerblom, M. Uhari, and E.M. Nuutinen. Regional differences in apolipoprotein E polymorphism in Finland. *Annals of medicine*, 23:61–66, 1991.
11. J. Li. On optimal rule discovery. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):460–471, 2006.
12. Y. Liang and A. Kelemen. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys*, 2:43–60, 2008.
13. P.A. Lotufo, U.C. Claudia, U.A. Ajani, C.H. Hennekens, and J.E. Manson. Male pattern baldness and coronary heart disease: the physicians' health study. *Archives Internal Medicine*, 160:165–171, 2000.
14. J.H. Moore, F.W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.
15. S. Morishita and J. Sese. Transversing itemset lattices with statistical metric pruning. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'00)*, pages 226–236. ACM Press, 2000.
16. S. Nijssen, T. Guns, and L. De Raedt. Correlated itemset mining in ROC space: a constraint programming approach. In *Proceedings the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 647–656. ACM Press, 2009.
17. S. Nijssen and J.N. Kok. Multi-class correlated pattern mining. In *Proceedings of the 4th International Workshop on Knowledge Discovery in Inductive Databases*, volume 3933 of *Lecture Notes in Computer Science*, pages 165–187. Springer, 2006.
18. G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.
19. D.R. Thiruvady and G.I. Webb. Mining negative rules using GRD. In *Advances in Knowledge Discovery and Data Mining, Proceedings of the 8th Pacific-Asia Conference (PAKDD 2004)*, volume 3056 of *Lecture Notes in Computer Science*, pages 161–165. Springer, 2004.
20. G.I. Webb. MagnumOpus software. `http://www.giwebb.com/index.html`. Retrieved 10.2. 2009.
21. G.I. Webb. Discovering significant rules. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, pages 434–443. ACM Press, 2006.
22. G.I. Webb and S. Zhang. K-optimal rule discovery. *Data Mining and Knowledge Discovery*, 10(1):39–79, 2005.
23. X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, 22(3):381–405, 2004.
24. F. Yates. Test of significance for 2 x 2 contingency tables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):426–463, 1984.