hyväksymispäivä arvosana

arvostelija

Relevantti informaatio koneoppimisessa

Ville H. Tuulos

Helsinki 29. maaliskuuta 2002

Älykkäiden järjestelmien tutkimusseminaari

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Sisältö

1 Prologi **1**

2 Similariteettimitat oppimisessa **2**

2.1 Piirteenirrotus 3

2.2 Similariteettimitat 5

3 Relevantti informaatio **7**

3.1 Pakkaus 8

3.2 Rate distortion theory 10

4 Information bottleneck **10**

4.1 Agglomerative information bottleneck 12

4.2 Multivariate information bottleneck 13

4.2.1 Multi-Information Principle 15

4.2.2 Variational Principle 16

5 Muita menetelmiä **17**

6 Epilogi **18**

Lähteet **19**

1 Prologi

Koneoppimisen tutkimuksen pitkä historia on jättänyt jälkeensä lukemattoman määrän julkaisuja. Näistä suurin osa käsittelee parametrien oppimista jossain yleisesti hyväksytyssä mallistruktuurissa. Huomattavasti vähemmän löytyy papereita, joissa esitellään uusia struktuureja tai niiden oppimista datasta. Näihin verrattuna ainoastaan häviävä määrä julkaisuja käsittelee yleisellä tasolla sitä, millainen data on otollista oppimisen kannalta.

Tämä työ on lyhyt katsaus perimmäiseen kysymykseen relevantin informaation olemuksesta. Aihe on kiinnostava huomattavan monelta näkökannalta, ulottaen vaikutuksena lähes kaikkeen, mikä liittyy "oppimiseen datasta". Luonnolliset teoreettiset puitteet käsittelylle löytyy informaatioteoriasta. Käy ilmi, että kommunikaatiokanavien ominaisuuksien kuvailuun luotu teoria voidaan laajentaa koskemaan myös kommunikaation sisältöä.

Työ nojautuu vuonna 1999 Tishbyn, Pereiran ja Bialekin esittelemään *Information Bottleneck* -menetelmään [TPB99]. Viimeisimpänä laajennoksena esitellään myös menetelmän monimuuttujaversio [Fri01]. Aihetta pohditaan ennen kaikkea luonnollisen kielen käsittelyn näkökulmasta. Tämän valinnan ei pitäisi rajoittaa havaintojen ja tulosten yleistyvyyttä muihinkin materiaaleihin, sillä useissa tapauksissa ongelma voidaan palauttaa luonnollisen kielen käsittelyongelmaksi. Lisäksi ongelmat, joihin luonnollisen kielen käsittelyssä törmätään, ovat pitkälti samoja kuin minkä tahansa muunkin reaalimaailmasta kerätyn aineiston kanssa. Menetelmien alkuperäiset englanninkieliset nimet ja fraasit on jätetty suomentamatta, jotta esityksen luettavuus suhteissa lähteisiin säilyisi.

Katsaus etenee seuraavasti: Seuraavassa kappaleessa luodaan katsaus ongelmiin perinteisessä koneoppimiskäsityksessä. Kolmas kappale käsittelee informaatioteoreettisten käsitteiden suhdetta relevantin informaation määrittelyyn. Tästä suora jatkumo on neljännen kappaleen katsaus Information Bottleneck -menetelmään, jota puolestaan viides kappale vertailee suhteessa muihin vastaaviin menetelmiin. Esityksen päättää lyhyt yhteenveto.

2 Similariteettimitat oppimisessa

Erityisesti useat ohjaamattoman oppimisen menetelmistä[1] perustuvat ajatukselle, että tunnetaan mitta, jolla yksittäisten syötteiden samankaltaisuutta voidaan mitata. Intuitio on luonnollinen, kun ryhmittelyä lähestytään seuraavalla ajatuksella: Oletetaan piirteenirrottaja $\Phi(\{\mathbf{R}\}) \rightarrow \{\mathbf{X}\}, \mathbf{X} \in \mathbb{R}^d$, joka kvantifioi syöteolioiden joukon $\{\mathbf{R}\}$ d-ulotteisen avaruuden vektorijoukoksi $\{\mathbf{X}\}$. Ryhmitellään saatu vektoriavaruus siten, että ryhmien sisällä etäisyydet ovat mahdollisimman pieniä ja ryhmien välillä mahdollisimman suuria. Tämä ajatus on perinteisimmän vektoriavaruuden käsittelyyn perustuvan ohjaamattoman oppimisen takana (k-Nearest Neighbor, k-means).

Samaan ajatukseen perustuu koko hierarkisten ryhmittelijöiden perhe, jotka alhaalta ylöspäin kokoavat puuta yhdistäen aina lähimpiä syötevektoreita muodostaen näin oksaryppäitä. Hieman tästä poikkeava vektorikvantisointiin perustuvan ryhmittelyn intuitio on, että syötevektoreista voidaan muodostaa huomattavasti pienempi joukko prototyyppivektoreita, joiden kohinaisiksi instansseiksi syötevektorit voidaan ryhmitellä – etsimällä jokaiselle syötevektorille lähin prototyyppivektori. Tähän koulukuntaan kuuluvat niinikään tunnetut menetelmät SOM (Self-Organizing Map) ja tämän sisarmenetelmä ohjatussa oppimisessa LVQ (Learning Vector Quantisation).

On ehkä merkillepantavaa, ettei yksikään edellä mainituista ryhmittelyparadigmois-

[1] Ohjaamaton oppiminen tulkitaan tässä synonyymiksi ryhmittelylle (klusteroinnille)

ta perustu tilastollisen oppimisteorian oletukseen, että syötemateriaali olisi jonkin (tuntemattoman) stokastisen prosessin tuottamaa ja tavoitteena olisi tämän prosessin mallintaminen. Similariteettitarkastelun ulkopuolelle jää siten probabilistisen ryhmittelyn menetelmät (esim. ääreelliset sekajakaumamallit).

2.1 Piirteenirrotus

Edellisessä kappaleessa esitelty piirteenirrottaja $\Phi(\{\mathbf{R}\})$ on mystinen. Mekanismi perustuu seuraavaan epäteoreettiseen intuitioon maailmasta: Olioiden ominaisuudet, piirteet, oletetaan kvantifioituviksi, olkoonkin niiden alkuperä kvalitatiivinen tai "luonnostaan" numeerinen. Jo kvantifiointivaihe kadottaa informaatiota - luonnollisen maailman oliot ovat jatkuvia, eikä niitä täydellisesti voi kuvata äärellisellä joukolla muuttujia. Pyrkimyksenä onkin löytää oliota parhaiten luonnehtivat merkittävät piirteet, jotka muodostavat olion representaation, piirrevektorin, piirreavaruuteen. Epäonnistunut piirteenirrotusprosessi tuottaa avaruuden, jossa olioiden keskinäiset suhteet eivät enään vastaa todellisuutta.

Kaikkien piirrekombinaatioiden evaluointi suhteessa oppimistuloksiin on selvästikin epäkäytännöllistä. On kuitenkin ilmeistä, ettei kaikkien piirteiden merkitys ole yhtä suuri oppimisongelmassa. Kun luonnollista kieltä ryhmitellään aihealueittain, ei sana 'ja' ole erityisen diskriminoiva piirre. Osa piirteistä voidaan siis tulkita *epärelevanteiksi* suhteessa oppimisongelmaan. Jäljelle jääneet tulkitaan *relevanteiksi*. Varsin usein tämän jaottelu tehdään adhoc-otteella tutkijan intuition perusteella. Kuitenkin harvoin tämä heikoin lenkki prosessissa joutuu syntipukiksi huonoista oppimistuloksista, vaan heikon laadun syyksi epäillään mallistruktuuria tai ongelmia parametrien oppimisessa (lokaali minimi tms.).

Automaattisia menetelmiä piirteidenvalintaan on kehitetty. Jos piirrekombinaatioiden määrä on edes jotenkin hallittavissa, voidaan sopivalla hakumenettelyllä käydä kombinaatiot läpi ja evaluoida tulokset suhteessa lopputulokseen. Piirteitä voidaan arvottaa ja valita sopivalla mitalla "sopivin" kombinaatio. Käytännössä koeteltuin menetelmä (tutkijan tekemän piirteiden adhoc-valinnan jälkeen) lienee kuitenkin sopivalla tavalla tehty piirreavaruuden dimension tiputtaminen, piirreavaruuden transformaatio.

Lineaarisia menetelmiä on useita, esim. PCA ja ICA (Principal- ja Independent Component Analysis), samoin epälineaarisia, kuten SOM. Teksiluokittelussa myös paljon käytetty menetelmä tähän on LSI (Latent semantic indexing). Intuitio tämän menettelyn takana on pudottamalla piirreavaruuden dimensiota sopivalla tavalla päästä käsiksi merkityksekkäisiin piirteisiin.

Alkuperäisessä WEBSOM -järjestelmässä [Kas98] hyödynnettiin itseorganisoivaa karttaa juuri tässä hengessä, muodostamalla kartta aluksi syötedokumenttien sanoista ja järjestelemällä dokumentit käyttäen hyväksi tämän sanakartan (joka siis on alkuperäisen piirreavaruuden epälineaarinen projektio) avulla muodostettuja histogrammeja. Kaksitasoisesta menetelmästä luovuttiin myöhemmin heikkojen tulosten vuoksi.

Epäilemättä piirteenirrotuksessa pyritään hapuilemaan kohti hyviä oppimistuloksia. Teoreettiset puitteet ovat kuitenkin sekä automaattisessa- että adhoc-piirteenirrotuksessa heikot. Pystytäänkö täyttämään edes luokittelijan tekemät (implisiittiset) oletukset avaruuden luonteesta? Koko ryhmittely perustuu ajatukseen tiheistä, koherenteista ryhmistä avaruudessa. Tuottaako valittu piirteenirrottaja tässä mielessä mielekästä dataa? K-means -ryhmittely olettaa ryhmien olevan pallomaisia. Entäpä voidaanko mielivaltainen piirteenirrottaja tulkita tilastollisessa mielessä sopivaksi stokastiseksi prosessiksi? Useat oppijoista ovat teoreettisesti universaaleja, MLP on universaali funktioapproksimaattori ja sekajakaumamallit toimivat universaaleina tiheysapproksimoijina ja täten ne toimivat "oikein" piirteenirrotusprosessista riippumatta. Kuitenkin tuloksia tarkasteleva ihminen on harvoin suvaitsevuudeltaan universaali; oppimistulosten tulisi yleensä vastata edes löyhimpiä ennakkohypoteeseja. Similariteettiperustaisten oppijoiden kannalta kriittinen kysymys onkin, vastaako kvantifioitujen piirrevektorien similariteettimitta ja ihmisen similariteettitulkinta vastaavista reaalimaailman olioista toisiaan.

2.2 Similariteettimitat

Metrinen tulkinta vektoriavaruudessa on ilmeinen, muttei ongelmaton. Yllättäen matalaulotteisen intuitiomme varassa tehdyt tulkinnat samankaltaisuudesta, etäisyydestä ja naapurustosta voivatkin pettää korkeaulotteisissa avaruuksissa. Ilmeinen tulkinta piste-etäisyyksistä on euklidinen. Kuitenkin tämä tulkinta voi olla kvalitatiivisessa ristiriidassa käytetyn piirteenirrotusprosessin kanssa. Herää perustavanlaatuinen kysymys, voiko ylipäätänsä geneeristä etäisyysmittaa olla olemassa data-analyysin tarpeisiin?

Koko ryhmittelykonseptin mielekkyys perustuu *klusterihypoteesiin*: relevantit oliot (dokumentit) ovat lähempänä toisiaan kuin epärelevantit. Kuitenkin usein yllättävät tulokset, mitä laajalti tutkittu ala on tuottanut, ovat antaneet aihetta paneutua tarkemmin hypoteesin oikeutukseen. Sopivilla, usein käytännössäkin törmättävillä, löyhillä oletuksilla käy ilmi, että avaruuden dimension kasvaessa läheisimmän ja etäisimmän pisteen suhde lähenee ykköstä eli kontrastit katoavat [Bey99]. Tuloksella on muutakin kuin vain teoreettista mielenkiintoa, sillä ilmiö sumentaa naapurustoja jo 15 ulottuvuudessa.

Euklidinen mitta on erikoistapaus yleisestä L_k normista, joka määritellään

$$L_k(x, y) := \sum_{i=1}^{d} (\|x^i - y^i\|^k)^{1/k}$$

kun $k = 2$ on mitta euklidinen. Jonkin verran euklidista mittaa parempaan erottelutarkkuuteen päästään korkeaulotteisissa avaruuksissa city-block -etäisyydellä, eli kun $k = 1$. Mielenkiintoiset empiiriset kokeet (teorian tuella) osoittavat lisäksi, että tulokset parantuvat entisestään käytettäessä fraktionaalisia mittoja $k \leq 1$. Kokeissa *k-means* ryhmittelyn luokittelutulos parani noin kymmenen prosenttia kun $k = 0.3$, euklidiseen mittaan $k = 2$ verrattuna [AHK01].

Tämä ei ole kuitenkaan ainoa ongelma. Euklidinen etäisyys tulkitsee kaikki komponentit yhtä painaviksi globaalisti. Kuitenkin globaali etäisyys (so. etäisyys kahden mielivaltaisen avaruuden pisteen välillä) voi olla mahdoton määritellä. Piirteenirrotuksessa kvantifioidut alunperin kvalitatiiviset ominaisuudet voivat olla mielekkäitä vertailla vain lokaalissa mielessä.

Vaikka teoreettisille tuloksille etäisyysmittojen käyttäytymisestä ei annettaisi paljoa painoa, on käytäntö hyvin vahvasti osoittanut, että ryhmittely (tai yleisemmin oppimistulos) on harvoin yksikäsitteinen. Ongelma ei ole käytetyissä menetelmissä. Ryhmittelyä tai mallintamista käytetään usein eksploratiivisessa hengessä hypoteesinmuodostukseen ja datan tulkintaan. Tästä asetelmasta ei voi olettaa syntyvän yksikäsitteisiä tuloksia.

Yksikäsitteisyyden illuusio tässä yhteydessä onkin syytä unohtaa ja kysyä ennemmin, voivatko tulokset olla mielivaltaisia. Kuten jo aiemmassa kappaleessa todettiin, menetelmät ovat usein universaaleja, mutta tulkitsija ei. On siis tarkasteltava ongelmaa tulkitsijan kannalta: mikä tieto on hänelle relevanttia? Sama dokumenttijoukko voidaan ryhmitellä helposti lähes mielivaltaisesti tulkintaa vaihtamalla – oleellinen kysymys onkin, vastaako ryhmittely annettua tulkintaa. Tämä muuttaa koko suhtautumista ryhmittelyyn tiedonlouhinnassa[2]: ryhmittely ei ole staattinen vaan

[2]Kysymys on mielenkiintoinen myös laajemmin oppimisen ja mallintamisen kannalta. Kuinka hyvin bayesiläisen mallintamisen priorit vastaavat haasteeseen? Tässä valossa myös käsite *Rein-*

muodostetaan dynaamisesti käyttäjän intentiota noudatellen. Kokemukset tästä lähestymistavasta ovat perinteisiäkin ryhmittelymenetelmiä käyttäen olleet lupaavia (esimerkiksi [HeP96]).

3 Relevantti informaatio

Edellisessä kappaleessa osoitettiin kaksi perustavanlaatuista ongelmaa perinteisessä vektoriavaruuksiin ja etäisyysmittoihin perustuvassa ryhmittelyssä tai koneoppimisessa yleensäkin. Piirteiden valinta ei ole käytännössä useinkaan automaattisesti mahdollista ja on manuaalisesti epävarmaa. Piirreavaruuden transformaatioihin perustuvat automaattiset menetelmät ovat vaihtelevasti perusteltuja ja niiden tuottamat tulokset eivät aina vastaa suinkaan sitä, mikä tulkitsijan kannalta olisi mielekästä. Samantyyppiset ongelmat vaivaavat etäisyyden mittaamista piirreavaruudessa, jossa lähekkäisten pisteiden tulisi olla jossain mielessä samankaltaisia. Perinteinen euklidinen mitta toimii epävarmasti korkea-ulotteisissa tapauksissa. Yksikäsitteisen similariteetin määrittäminen, joka toimisi tapauksessa kuin tapauksessa, tuntuu olevan mahdotonta ja valinta on usein heuristinen. Similariteetin määrityksen tulisi olla kontekstisidonnaista. Selvästikin olisi tilaus formalismille, joka kykenisi vastaamaan geneerisesti edellä esiteltyihin ongelmiin.

Seuraavassa konstruoidaan informaatioteoreettinen formalismi, joka kykenee antamaan vastauksen näihin kysymyksiin noudatellen Information Bottleneck -menetelmän määrittelyä [TPB99]. Formalismin ydin on relevantin informaation määritelmä, joka laajentaa Shannonin alunperin vain informaation muotoa käsittelevän teorian koskemaan myös sen sisältöä.

Joitan perusedellytyksiä määritelmälle voidaan koota aiemmasta tarkastelusta: Mää-

forcement learning vaikuttaa kiinnostavalta.

ritelmä ei saa tehdä käytännössä rajoittavia oletuksia käsiteltävästä datasta (geneerisyys). Määritelmän oikeutus tulee olla teoreettisesti perusteltavissa (uskottavuus). Määritelmän avulla tulee voida yksikäsitteisesti käytännöllisin menetelmin määrittää piirteen (satunnaismuuttujan) sisältämä informaatio suhteessa toiseen satunnaismuuttujaan (relevanssi). Informaatioteoreettinen käsite *mutual information* $I(X;Y) := H(X) - H(X|Y)$ täyttää ehdot. Mutual information (MI) kertoo, kuinka paljon muuttuja X sisältää informaatiota muuttujasta Y. Muuttujan X relevanssi voidaa täten määrittää relevanssimuuttujan Y avulla. Lisämotivaatio MI:n käyttöön oppimisongelmissa saadaan Fanon epäyhtälöstä [FiP98]

$$P(C \neq \bar{C}) \geq \frac{H(C|X) - 1}{\log N} = \frac{H(C) - I(C;X) - 1}{\log N}$$

Epäyhtälö kertoo seuraavaa: C on diskreetisti jakautunut luokkamuuttuja ja $\bar{C}$ on C:n estimaatti X havaitsemisen jälkeen. N on C:n kuvaaminen luokkien lukumäärä. Epäyhtälö ilmoittaa, että luokitteluvirheen todennäköisyys $P(C \neq \bar{C})$ on alhaalta rajoitettu ja pienimmillään silloin kun X:n sisältämä informaatio C:stä [3] $I(C;X)$ on mahdollisimman suuri. Eli pyrittäessä mahdollisimman hyvään luokitteluun, on mielekästä pyrkiä maksimoimaan MI:a.

3.1 Pakkaus

Häviöllinen pakkaus on toinen informaatioteorian käsite, joka on yllättävänkin läheisessä suhteessa piirteidenvalintaan ja relevanttiin informaatioon. Tilanne on ensisilmäyksellä kiero. Toisaalta sekä ihmis- että koneoppimien on mahdollista vain redudantin informaation avulla (hauki on kala). Toisaalta häviöllisessä pakkauksessa halutaan päästä eroon redudanssista entropian yli. Piirteenirrotuksessa pakkausta usein käytetään pudottamaan alkuperäisen avaruuden dimensiota ja vähentämään kohinaa.

[3] Symmetrian nojalla, sillä $I(X;Y) := H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Tällä ajatuksella on myös dokumenttiluokittelussa hyödynnetty termiryhmittelyä. Ajatus on hyvä: ryhmitellään samoissa konteksteissa esiintyvät termit, jolloin muodostuneet ryhmät kuvaavat yksittäisiä sanoja paremmin dokumentin teemoja. Menetelmän avulla päästään eroon termivektorien harvuudesta ja ratkaistaan elegantisti sanojen monitulkintaisuusongelma. Kysymys on häviöllisestä pakkauksesta: relevantti informaatio dokumenttiluokittelun kannalta on dokumenttien teemat, eivät yksittäiset sanat. Sanat voidaan siis korvata huoletta viittauksilla vastaaviin ryhmiin, mikä usein tiputtaa rajusti avaruuden dimensiota. Kuitenkaan menetelmä ei ole de facto -asemassa ja saadut tulokset vaihtelevat suuresti. Syynä ovat usein luvussa 2. mainitut ongelmat. Häviöllistä pakkausta on pyritty implisiittisesti hyödyntämään, mutta ilman geneeristä menetelmää tai tukevaa pohjateoriaa – seurauksena lopputulokset ovat vettyneet.

Seuraavassa esitetään korkean tason muotoilu edellä esitetylle ajatukselle piirteiden pakkauksesta. Merkitään X:llä alkuperäisiä piirrevektoreita ja $p(x)$:llä vastaavia todennäköisyyksiä. Jokaiselle $x \in X$ etsitään vastaava kuva pakattujen representaatioiden (koodisanojen) joukossa $\bar{x} \in \bar{X}$, esimerkiksi sopivalla stokastisella kuvauksella. Tämä vastaa X:n pehmeää ryhmittelyä, jossa alkuperäiset vektorit korvataan koodisanoilla $\bar{x} \in \bar{X}$ todennäköisyydellä $p(\bar{x}|x)$. Koodisanan kokonaistodennäköisyydeksi saadaan:

$$p(\bar{x}) = \sum_{x \in X} p(x) p(\bar{x}|x)$$

Keskimäärin koodisanan informaatiosisältö saadaan ehdollisen entropian avulla ja on $2^{H(X|\bar{X})}$. Millä mitataan hyvää pakkausta? Luonnollinen mitta on keskimääräinen bittien määrä koodisanassa, joka tarvitaan, jotta viesti säilyy ymmärrettävänä. Tämä saadaan suhteesta $2^{H(X)}/2^{H(X|\bar{X})} = 2^{I(X;\bar{X})}$ eli suoraan MI mitta, kuinka paljon $\bar{X}$ säilyttää tietoa X:stä

$$I(X; \bar{X}) = \sum_{x \in X} \sum_{\bar{x} \in \bar{X}} p(x, \bar{x}) \log \frac{p(\bar{x}|x)}{p(\bar{x})}.$$

Kuitenkin selvästi tämä informaation määrä saadaan mielivaltaisen pieniksi yksityiskohtia kadottamalla, joten lisärajoitukset ovat tarpeen.

3.2 Rate distortion theory

Selvästikään pakkauksesta ei ole hyötyä jos se kadottaa liikaa informaatiota alkuperäisestä muuttujasta. Toisaalta yksikäsitteinen tavoite on päästä käsiksi vain relevanttiin informaatioon ja muulla muuttujan sisältämällä informaatiolla ei ole merkitystä. Tarvitaan siis menetelmä, joka kertoo säilytettävän informaation ja sallitun hävikin optimaalisen suhteen. Shannonin ja Kolmogorovin *rate distortion theory* tarjoaa ongelmaan tyylikkään ratkaisun.

Tarvitaan hävikkiä kuvaava funktio (distorsion function) $d : X \times \bar{X} \to \mathbb{R}^+$, joka saa pieniä arvoja hävikin ollessa pieni. Tämän mitan avulla voidaan edellisessä kappaleessa lasketun pakkauksen hävikki laskea

$$G_{p(x,\bar{x})} = \sum_{x \in X} \sum_{\bar{x} \in \bar{X}} p(x, \bar{x}) d(x, \bar{x}).$$

Mitä maltillisempaan pakkaukseen tyydytään sitä pienempänä pysyy hävikki. Tätä suhdetta kuvaa *rate distortion function* R(D), joka kertoo suurimman saavutettavan pakkauksen sallitussa hävikin puitteissa:

$$R(D) := \min_{p(\bar{x}|x): G_{p(x,\bar{x})} \leq D} I(X; \bar{X}).$$

Blahut-Arimoton algoritmilla voidaan $R(D)$ numeerisesti ratkaista.

4 Information bottleneck

Edellä esitellyt teemat, pakkaus ja relevantin informaation informaatioteoreettinen muotoilu, antavat formaalin muotoilun luvussa 2. esitettyihin ongelmiin. Yksinään ei kummastakaan ole kokonaisongelman ratkaisuksi, sillä monet merkittävät kysymykset jäivät avoimiksi. Kuinka kuvaus alkuperäisiltä piirrevektoreilta koodisanojen joukkoon $\bar{x} \in \bar{X}$ käytännössä tehdään? Tämähän oli ryhmittelyn alkuperäinen ongelma. Kuinka MI:a voidaan käytännössä hyödyntää?

Luonnollinen ratkaisu on yhdistää nämä kaksi konseptia. Tuloksena saadaan menetelmä, jota Tishby, Pereira ja Bialek kutsuvat Information Bottleneck:ksi. Koska yksikäsitteisen oikeaa hävikkimittaa ei yleensä ole, tehdään pakkaus siten, että pakattava muuttuja X kadottaa mahdollisimman vähän relevanttia informaatiota, joka mitataan suhteessa muuttujaan Y. Tämä on melko suora laajennos edellisen kappaleen hävikkifunktioista, mutta oleellista on juuri toisen "apumuuttujan" Y käyttöönotto, jonka avulla "oikea" hävikkifunktio voidaan määrätä. Halutaan siis maksimoida $I(\bar{X};Y)$ ja minimoida $I(\bar{X};X)$. Tämä luonnollisesti edellyttää, että X ja Y eivät ole riippumattomia. Lisäksi yhteistodennäköisyysjakauma $p(x,y)$ oletetaan tunnetuksi.

Optimointiongelma muotoillaan seuraavasti:

$$I(\bar{X};Y) = \sum_y \sum_{\bar{x}} p(y,\bar{x}) \log \frac{p(y,\bar{x})}{p(y)p(\bar{x})}$$

Mikä triviaalisti on tietenkin pienempi tai yhtäsuuri kuin alkuperäinen $I(X;Y)$. Tämä kuvaa myös ajatusta pullonkaulasta: X ikäänkuin sullotaan ahtaan pullonkaulan läpi, joka säilyttää X:stä vain Y:stä kertovan informaation. Jokseenkin kanoninen tapa ratkaista ääriarvoongelma, johon sisältyy rajoituksia, ovat Lagrangen kertoimet. Tällä menetelmällä ongelma voidaan muotoilla seuraavasti:

$$\mathcal{L}[p(\bar{x}|x)] = I(\bar{X};X) - \beta I(\bar{X};Y)$$

Tässä β on Lagrangen kerroin määräten pakkauksen ja säilytettävän informaation suhteen; kun $\beta = 0$ katoaa Y:n merkitys ja pakkaus on mielivaltainen ja kun $\beta \to \infty$ säilytettävien yksityiskohtien määrä kasvaa. Koko Information Bottleneckin mielenkiintoisuutta lisää huomattavasti se, että edelliselle minimointiongelmalle löytyy eksakti ratkaisu, joka todistuksineen löytyy lähteestä [TPB99].

Vielä ei ole otettu kantaa ryhmittelymuuttujan $\bar{X}$ kardinaliteettiin eikä rakenteeseen (so. ryhmittelyyn). Alkuperäisessä Information Bottleneck menetelmässä ryhmittely perustui *deterministic annealing* -tyyppiseen ryhmittelyyn, jossa järjestelmän vapaan energian määräksi tulkitaan[4] $T = \beta^{-1}$. Jokaista $\bar{X}$:n kardinaliteettia vastaa oma käyränsä ja tämän käyräperheen parittaiset käyrät eroavat jollakin β:n arvolla. Näitä "tilasiirtymiä" hyväksikäyttäen voidaan muodostaa hierarkinen ryhmittely ylhäältä alaspäin. Menetelmä esitellään tarkemmin lähteessä [PTL93].

Mitä Information Bottleneck siis tekee? Yleisesti ottaen Information Bottleneck tarjoaa formaalit puitteet useisiin tiedonkäsittelyongelmiin suodatuksesta koneoppimiseen. Information Bottleneck ryhmittelee alkuperäisen piirreinformaation käyttäen hyväksi annettua "mittaa" relevanssille. Menetelmä ei tee oletuksia alkuperäisen datan jakaumasta eikä lähteestä (kuten implisiittisesti probabilistinen ryhmittely) eikä perustu mielivaltaiseen vektorointiin ja vektoriavaruuden (euklidiseen) tulkintaan. Laskenta perustuu tältä osin vain yhteistodennäköisyysjakaumaan $p(x, y)$, joka esimerkiksi dokumenttiluokittelun tapauksessa voidaan laskea sanojen X ja dokumenttien Y yhteisesiintymämatriisista.

4.1 Agglomerative information bottleneck

Alkuperäisen Information Bottleneckin ylhäältä alaspäin etenevä ryhmittely ei ole ongelmaton. Tilasiirtymien tunnistaminen ei ole helppoa, eivätkä tulokset ole deterministisiä. Vaihtoehtoiseksi ratkaisuksi Slonim ja Tishby esittävät [SlT99] perinteistä alhaalta ylöspäin etenevää sulauttavaa hierarkista ryhmittelyä. Yleisesti sulauttava

[4] Tämä on luonnollista, sillä $T \to 0$ kun $\beta \to \infty$ eli vapaa energia vähenee kun ryhmittely "kovenee".

hierarkinen ryhmittely (*agglomerative hierarchial clustering*) perustuu seuraavaan ajatukseen: Lähtötilanteessa tulkitaan jokainen X:n alkio omaksi ryhmäkseen, jotka muodostavat puun lehdet. Seuraavassa vaiheessa yhdistetään toisiaan lähimpänä olevat alkiot näiden yhteiseksi isäsolmuksi. Menettelyä toistetaan kunnes jäljellä on vain yksi ryhmä (puun juuri)[5]. Algoritmi on ahne ja tuottaa ainakin lokaalisti optimaalisia ratkaisuja. Lisäksi ryhmittely voidaan tehdä ilman vaikeaa vapaiden parametrien määrittämistä.

Information Bottleneckin tapauksessa on solmujen etäisyysmitan valinta luonnollinen. Sulautukset valitaan siten, että paikallisesti hävitetään mahdollisimman vähän yhteistä informaatiota $I(\bar{X};Y)$. Menetelmän suhdetta optimaaliseen ryhmittelyyn on vaikea osoittaa, mutta ilmeisesti kun relevanssimuuttuja Y on binäärinen, on ryhmittely optimaalinen [SlT99]. Kuitenkin parhaita tuloksia saataneen ryhmittelyssä, jossa alkuperäinen *deterministic annealing* ja kova hierarkinen ryhmittely yhdistetään. Puu muodostetaan alhaalta ylöspäin ja lopuksi pehmennetään alkuperäisellä menetelmällä ylhäältä alaspäin, jolloin tilasiirtymät tulevat automaattisesti määriteltyä.

Kuten aina hierarkisen ryhmittelyn kanssa, on suurin ongelma menetelmän tehokkuudessa. Koska jokainen alkio tulkitaan aluksi omaksi ryhmäkseen, on menetelmä huonosti skaalautuva suurille aineistoille. Pahimman tapauksen aikavaatimus on $O(|X|^3)$.

4.2 Multivariate information bottleneck

Aina ei relevanssimuuttujaa Y ole helppo määrittää. On mahdollista, että X:n relevanssi määrittyy useamman kuin yhden muuttujan kautta. Toisaalta tavoitteet

[5]Vertaa algoritmia Huffman-koodaukseen!

ryhmmittelyn suhteen voivat olla moninaisemmat. Esimerkkinä olkoon dokumenttiryhmittely, jossa dokumenttitietoa Y käyttäen muodostetaan termiryhmät $\bar{X}$ ja puolestaan tätä tietoa käytetään hyväksi dokumenttien ryhmittelyssä. Jotta Information Bottleneck -menetelmää voitaisiin tehokkaasti soveltaa tämänkaltaisissa tilanteissa, on menetelmää laajennettava kuvaamaan useamman muuttujan välisiä suhteita. Seuraavassa esitellään Friedman, Mosenzon, Slonim ja Tishbyn Multivariate information bottleneck [Fri01], joka vastaa tähän tarpeeseen.

Bayes-verkot ovat varsin käyttökelpoinen formalismi monimuuttujaisten todennäköisyysjakaumien kuvailuun ja käsittelyyn. Bayes-verkko G on suuntaamanton syklitön verkko, jonka solmuja ovat satunnaismuuttujat $X_1, ..., X_n$. Muuttujan X_i vanhempia verkossa G merkitään joukolla $Pa^G_{X_i}$. Jakauma P on konsistentti verkon G kanssa jos se voidaan muotoilla seuraavasti:

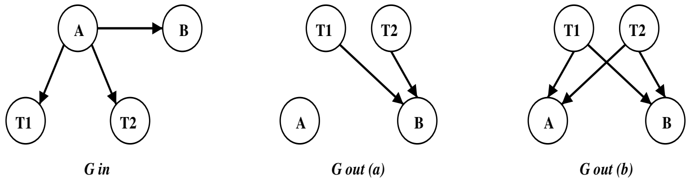
$$P(X_i, ..., X_n) = \prod_i P(X_i | Pa^G_{X_i})$$

Jos P on konsistentti verkon G kanssa, voidaan MI:n monimuuttujaversio muotoilla seuraavasti:

$$I^G := I(X_i, ..., X_n) = \sum_i I(X_i; Pa^G_{X_i})$$

Alkuperäinsen koodikirjan $\bar{X}$ sijaan tarvitaan pakattavien ryhmittelymuuttujien joukko $\mathbf{T} = \{T_1, ..., T_k\}$. Ryhmittelumuuttujien suhde havaittujen muuttujien joukkoon $\mathbf{X} = \{X_i, ..., X_n\}$ määritellään Bayes-verkolla G_{in} sisältäen solmut $\mathbf{X} \cup \mathbf{T}$. Ryhmittelymuuttujat $\mathbf{T}$ tulkitaan isäsolmujen funktioksi (kuten yhden muuttujan tapauksessakin), joten rajoitutaan tarkastelemaan verkkoja, jossa muuttujat $\mathbf{T}$ ovat lehtiä. Ryhmittelyosuus voidaan monimuuttujaversiossakin tehdä joko pehmeällä "jäähdytyksellä" tai kovalla hierarkisella menetelmällä.

Bayes-verkko G_{in} kuvaa ryhmittelymuuttujien suhdetta alkuperäiseen muuttujajoukkoon. Yhden muuttujan tapauksessa ryhmittelymuuttujan muodostaminen oli suoraviivaista. Monimuuttujaisessa tapauksessa tavoitteet eivät ole yhtä selkeitä.



Kuva 1: Esimerkki syöte- ja tavoiteverkoista

Lähtötilannetta ja tavoitteita kuvataan verkoilla, joista esimerkki kuvassa 1. Päämäärään voidaan päästä eri tavoin. Friedman et al. muotoilivat kaksi hieman erilaista lähestymistapaa.

4.2.1 Multi-Information Principle

Ensimmäinen tapa on alkuperäisen Information Bottleneckin suoraviivainen laajennos. Samoin kuin alkuperäisessä menetelmässä, halutaan **T**:n säilyttävän mahdollisimman paljon informaatiota relevanssimuuttujasta. Halutaan siis minimoida informaatiota $I^{G_{in}}$, mikä tekee muuttujista mahdollisimman riippumattomia. Samalla halutaan maksimoida informaatiota $I^{G_{out}}$, koska tämän voidaan tulkita kertovan, kuinka paljon informaatiota **T** säilyttää ennustettavista muuttujistaan.

Kuva 1. antaa esimerkin tilanteesta. Verkko G_{in} osoittaa muodostettavien ryhmittelymuuttujien T_1 ja T_2 riippuvan muuttujasta A. Halutaan, että T_1 ja T_2 irrottavat A:sta mahdollisimman paljon informaatiota B:n ennustamiseen, mitä tilannetta verkko $G_{out}(a)$ kuvaa. Tässä tilanteessa minimoidaan $I^{G_{in}} = I(A;B) + I(A;T_2) + I(A;T_1)$ ja maksimoidaan $I^{G_{out}} = I(B;T_1,T_2)$. Selvästikin oletetaan, että T_1 ja T_2 yhdessä pakkaavat A:n ja ennustavat B:tä riippumattomasti.

4.2.2 Variational Principle

Toinen tapa ei perustu enää informaation säilyttämiseen relevanssimuuttujista. Halutaankin säilyttää mahdollisimman paljon informaatiota annetusta kohdeverkosta. Ajatuksena on, että pakattujen muuttujien joukko $\mathbf{T}$ kuvaisi mahdollisimman hyvin koko kohdeverkon yhteistodennäköisyysjakaumaa. Kohdeverkon struktuuri on siten sellainen, jossa ennustettavat muuttujat riippuvat ainoastaan joukosta $\mathbf{T}$. Ongelmaksi jää konstruoida $P(T|A)$ siten, että nämä riippumattomuusoletukset toteutuvat. Tätä tulosta voidaan approksimoida minimoimalla jakaumien P ja verkon G_{out} kanssa konsistentin jakauman "etäisyyttä", joka saadaan tutusta Kullback-Leibler divergenssistä $D(P \parallel G_{out})$. Minimoitava Lagrangen funktio on täten:

$$\mathcal{L}' = I^{G_{in}} + \gamma D(P \parallel G_{out})$$

Kuvassa 1. verkko $G_{out}(b)$ kuvaa kohdeverkkoa. Informaatio ryhmittelymuuttujien T_1 ja T_2 välillä on minimissään (riippumattomia) ja maksimaalinen verkon B ennustamisen kannalta. Halutaan siis minimoida:

$$\mathcal{L}' = I(A;T_1) + I(A;T_2) + \gamma(I(T_1;T_2) - I(B;T_1,T_2))$$

Yleinen monimuuttujaversio mahdollistaa monimutkaistenkin ongelma-asettelujen muotuilun varioimalla verkkoja G_{in} ja G_{out}. Menetelmä tarjotaa varsin geneeriset formaalit puitteet, jotka myös mahdollistavat uuden muotuilun vanhoille data-analyysi -menetelmille, kuten ICA:lle [Fri01]. Artikkelissaan Friedman et al. antavat menetelmän soveltamisesta muutamia esimerkkejä.

5 Muita menetelmiä

Mutual Information on optimaalinen mitta piirreavaruuden transformaatioille. Useissa viimeaikaisissa tutkimuksissa onkin pyritty löytämään informaatioteoreettinen formalisointi menetelmille, jotka tavalla tai toisella pyrkivät pudottamaan piirreavaruuden dimensiota. Lähes aina MI on mukana tavalla tai toisella. Toinen keskeinen mielenkiinnon kohde on hankkiutuminen eroon heuristisesti valituista similariteettimitoista. Menetelmät muistuttavat suuresti toisiaan, sillä pohjalta löytyvät samat stokastiset ja informaatioteoreettiset formalismit. Tätä voidaan pitää pelkästään positiivisena, sillä luja yhteinen pohja mahdollistaa tiedon kumuloitumisen.

Information Bottleneck -menetelmän formaali pohja on luja ja itse menetelmä suhteellisen suoraviivainen. Parannettavaa löytyy yksityiskohdista (jotka kylläkin voivat olla perustavanlaatuisia) ja käytetyistä algoritmivalinnoista. Ryhmiteltävän muuttujan X ja relevanssimuuttujan Y yhteistodennäköisyysjakauma $p(x, y)$ oletetaan tunnetuksi. Käytännön sovelluksen kannalta tämä merkitsee, että jos muuttuja saa ennaltanäkemättömän arvon, ei ryhmää voida näytteelle määrätä. Tähän ongelmaan ottaa kantaa Kasken *Learning Metrics* [Kas01] tarjoten samalla visualisoinnissa usein käyttökelpoiselle itseorganisoivalle kartalle (SOM) luontevan informaatioteoreettisen etäisyysmitan. Menetelmä on muotoiltu vektorikvantisoinnin hengessä siinä missä Information Bottleneck perustuu häviöllisen pakkauksen teoriaan. Konseptit ovat kuitenkin hyvin läheisiä. Lisäksi menetelmä toimii jatkuva-arvoisille muuttujille toisin kuin Information Bottleneck, joka edellyttää jatkuva-arvoisten muuttujien diskretointia.

Käytännön sovellusten kannalta laskennallinen raskaus ja skaalautuvuuskysymykset voivat olla merkittäviä. Information Bottleneck ei tässä suhteessa ole erityisen kiitollinen, sillä menetelmän käyttämä hierarkinen ryhmittely (joko lehdistä juureen tai juuresta lehtiin) on raskas ajallisesti. Samoin eräät menetelmän käyttämät iteratiiviset algorimit ovat kohtuullisen raskaita. Torkkola on korvannut piirreavaruuksien transformaatioita käsittelevässä tutkimuksissaan [ToC00] Shannonin entropian Renyi'n kvadraattisella entropialla (diskreeteissä tapauksissa $-\log \sum_c p(c)^2$), joka on laskennallisesti kevyempi. Olisi mielenkiintoista nähdä, kuinka Renyi'n entropia istuisi Information Bottleneck -formalismiin.

6 Epilogi

Piirteiden valinta on vaikeaa; tapahtuu kombinatorinen räjähdys. On turvauduttava piirretransformaatioihin, mikä usein rinnastuu häviölliseen pakkaukseen. Euklidinen avaruus käyttäytyy huonosti korkeissa dimensioissa. Yksikäsitteistä similariteettimittaa on eksploratiivisessa data-analyysissa mahdoton muodostaa. Similariteetti on aina kontekstisidonnaista. Tämä ei pidä paikkaansa vain data-analyysissä, vaan myös datan generoivassa reaalimaailmassa. Hypoteeseista ja prioreista vapaa eksploratiivinen data-analyysi on illuusio – samoin ohjaamaton oppiminen. Ohjaus voi olla vain enemmän tai vähemmän näkyvää. Tämäkin pitää paikkaansa myös reaalimaailmalle. On parasta antaa ohjakset eksplisiittisesti tulkitsijalle ja antaa hänelle mahdollisuus määrätä mikä on relevanttia.

Information Bottleneck on käsittelyformalismi informaatiolle, joka on peräisin sumeasta ja kontekstisidonnaisesta maailmasta. Eksaktilta pohjalta on sumeuteen mahdollista tarttua. Mutual Information tarjoaa formalismin subjektiivisen relevanssin määrittelyyn ja käsittelyyn. Menetelmän monimuuttujaversio yhdistää informaatioteoreettisen pohjaan monimutkaisten ilmiöiden kuvailussa havainnolliset Bayes-verkot. Kombinaatio on kuvausvoimaltaan suuri, mutta iältään nuori. Vie aikansa ennen kuin tulosten käyttökelpoisuus ymmärretään kokonaisuudessaan.

Lähteet

AHK01 Aggarwal C., Hinneburg A., Keim A., On the surprising behavior of distance metrics in high dimensional space, *Proc. 8th International Conference on Database Theory*, Lontoo, Iso-Britannia, 2001, 420-434.

Bey99 Beyer K. et al., When is nearest neighbors meaningful, *Proc. International Conference on Database Theory*, 1999.

FiP98 Fischer J.W III, Principe J.C, A methodology for information theoretic feature extraction. *Proc. of IEEE World Congress on Computational Intelligence*, Anchorage, USA, 1998, 1712-1716.

Fri01 Friedman N. et al., Multivariate Information Bottleneck. *Proc. 17th Conference on Uncertainity in Artificial Intelligence*, Seattle, USA, 2001, 152-161.

HeP96 Hearst M., Pedersen J., Reexamining the cluster hypothesis: Scatter/Gather on retrieval results, *Proc. 19th Annual International ACM/SIGIR Conference*, Zurich, Sveitsi, 1996.

HAK00 Hinneburg A., Aggarwal C., Keim A., What is the nearest neighbor in high dimensional spaces?, *Proc. 26th International Conference on Very Large Databases*, Kairo, Egypti, 2000, 506-515.

Kas98 Kaski S. et al, Websom - self-organizing maps of document collections, *Neurocomputing*, 21, 1998, 101-117.

Kas01 Kaski S., Learning metrics for exploratory data analysis. *Neural Networks for Signal Processing XI, Proc. IEEE Signal Processing Society Workshop*, Toim. D. Miller, T. Adali, J. Larsen, M. Van Hulle ja S. Douglas, 2001, IEEE, New York, 53-62.

PTL93 Pereira F., Tishby N., Lee L., Distributional clustering of English words. *30th Annual Meeting of the Association for Computational Linguistics*, 1993, 183-190.

SlT99 Slonim N., Tishby N., Agglomerative Information Bottleneck. *Advances in Neural Information Processing Systems*, 12, 1999, 617-623.

TPB99 Tishby N., Pereira F., Bialek W., The Information Bottleneck Method. *Proc. 37th Allerton Conference on Communication and Computation*, 1999.

ToC00 Torkkola K., Campbell W., Mutual information in learning feature transformations *Proc. 17th International Conference on Machine Learning*, Stanford, USA, 2000.