# Bayesilaiset hybridiverkot

**Arto Klami**

**21. marraskuuta 2001**

Perustuu S.Montin ja G.F.Cooperin artikkeliin Learning Hybrid Bayesian Networks from Data teoksessa Learning in Graphical Models

## Sisältö

- Hybridiverkot - Bayesilaisia uskomusverkkoja diskreeteillä ja jatkuvilla muuttujilla
- Hybridiverkon rakenteen valinta
- Tiheysmallinnus neuroverkoilla
- Koetulokset
- Yhteenveto

## Hybridiverkko

- Perinteisesti bayesverkoissa muuttujat ovat joko diskreettejä tai normaalijakautuneita
- Diskreetissä tapauksessa todennäköisyydet esitetään todennäköisyystaulujen avulla ja normaalijakautuneet todennäköisyydet voidaan esittää keskiarvon ja keskihajonnan avulla
- Hybridiverkossa voi olla sekä diskreettejä että jatkuvia muuttujia samanaikaisesti ja lisäksi jatkuvat jakaumat voivat olla mielivaltaisen muotoisia. Mielivaltaiset jakaumat voidaan esittää esimerkiksi neuroverkkojen avulla.

## Verkon rakenteen valinta

- Bayesilaisessa analyysissa tulokset saadaan keskiarvoistamalla yli kaikkien mahdollisten verkkorakenteiden
- Käytännössä joudutaan kuitenkin valitsemaan paras yksittäinen rakenne
- Tutkitaan mallin uskottavuutta annetulla datalla ja valitaan uskottavin verkko

$$P(B_S|D) = P(B_S, D)/P(D) \propto P(B_S, D)$$

- Yleensä oletetaan että näytteet ovat samasta jakaumasta eikä datassa ole puuttuvia komponentteja

# Diskreetin bayesverkon rakenteen valinta

- Oletetaan globaali ja lokaali parametrien riippumattomuus eli eri muuttujien jakaumat ovat toisistaan riippumattomia ja lisäksi yhden muuttujan vanhempien eri realisaatioihin liittyvät jakaumat ovat riippumattomia toisistaan
- Tällöin valintakriteeri voidaan faktoroida muotoon

$$P(B_S|D) \propto P(B_S, D) = P(B_S) \prod_i^n s(x_i, \pi_i, D)$$

- Artikkelissa valittu diskreetille verkolle pisteytykseksi Cooper-Herskovitsin mitta

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(r_i)}{\Gamma(N_{ij} + r_i)} \prod_{k=1}^{r_i} \Gamma(N_{ijk})$$

# Verkon rakenteen valinta neuroverkkopisteytyksellä - 1

- Määrittelemällä uusi pisteytysmenetelmä päästään eroon parametrien lokaalista riippumattomuusoletuksesta
- Merkitään datan $l-1$:htä ensimmäistä näytettä $D_l$:llä ja suluissa olevalla yläindeksillä $l$:ttä arvoa
- Oletetaan faktoroituva priori $P(B_S, D) = \prod_i P(\pi_i)$ jolloin saadaan pisteytys

$$P(B_S, D) = \prod_{i=1}^{n} [P(\pi_i) \prod_{l=1}^{m} P(x_i^{(l)} | \pi_i^{(l)}, D_l, B_S)] = \prod_{i=1}^{n} s(x_i, \pi_i, D)$$

- Tämä vastaa sitä että edellisten näytteiden perusteella ennustetaan uuden näytteen todennäköisyys

## Verkon valinta neuroverkkopisteytyksellä - 2

- Termit $P(x_i|\pi_i, D_l, B_S)$ pitäisi laskea integroimalla kaikkien parametrien yli

$$P(x_i|\pi_i, D_l, B_S) = \int_{\theta_i} P(x_i|\pi_i, \theta_i, B_S)P(\theta_i|D_l, B_S)d\theta_i$$

- Yleensä suljettua ratkaisua ei voida laskea joten käytetään MAP-estimaattia
- Toisaalta MAP-estimaatti on asymptoottisesti ekvivalentti ML-estimaatin kanssa kaikilla positiivisilla prioreilla. Käytetään siis ML-estimaattia.
- Estimaatit saadaan neuroverkon, tässä tapauksessa monikerrosperseptronin, avulla

## Verkon valinta neuroverkkopisteytyksellä - 3

- Joudutaan laskemaan muuttujien määrä kertaa opetusnäytteiden määrän verran ML-estimaatteja
- Laskemista voidaan vähentää muodostamalla ML-estimaatti $t$:n näytteen välein ja olettamalla että se ei muutu estimointien välillä
- Voidaan myös antaa $t$:n kasvaa datan määrän mukana. Mitä enemmän dataa on käytetty nykyisten estimaattien muodostamiseen sitä enemmän näytteitä tarvitaan estimaatin muuttamiseksi.
- Voidaan esimerkiksi käyttää muotoa

$$t = \lceil \lambda l \rceil ,$$

jossa $l$ on tähän mennessä käsiteltyjen näytteiden määrä ja $0 < \lambda \leq 1$.

## Tiheysestimaattori diskreeteille muuttujille

- Diskreetit muuttujat mallinnetaan käyttämällä softmax-mallia
- Solmulla $x_i$ on $|\pi_i^c|$ jatkuvaa vanhempaa ja jokaisella diskreetillä vanhemmalla $x_j \in \pi_i^d$ on $r_j$ mahdollista arvoa
- Mallinnetaan todennäköisyys MLP:llä jossa on $|\pi_i^c| + \sum_{j \in \pi_i^d}(r_j - 1)$ sisääntuloa ja $r_i$ ulostuloa. $f_k(\pi_i)$ on lineaarinen funktio $k$:nnesta ulostulosta.

$$P(x_i = v_k|\pi_i) = \frac{e^{f_k(\pi_i)}}{\sum_{j=1}^{r_i} e^{f_j(\pi_i)}}$$

- Neliövirhettä minimoimalla tällainen verkko johtaa bayesilaisiin posterioritodennäköisyyksiin

## Tiheysestimaattori jatkuville muuttujille - 1

- Jatkuville muuttujille käytetään Bishopin esittelemää sekoitetiheysverkkoa (Mixture Density Network)
- Todennäköisyys mallinnetaan sekoitemallilla jossa on $K$ gaussista komponenttia

$$P(x_i|\pi_i) = \sum_{k=1}^{K} \alpha_k(\pi_i)\phi_k(x_i|\pi_i)$$

$$\phi_k(x_i|\pi_i) = c\exp\left(-\frac{(x_i - \mu_k(\pi_i))^2}{2\sigma_k(\pi_i)^2}\right)$$

## Tiheysestimaattori jatkuville muuttujille - 2

- Kaikkia mallin parametreja kontrolloidaan monikerrosperseptronin ulostuloilla. $K$:lle komponentille tarvitaan jokaiselle keskiarvo, varianssi ja painokerroin $\alpha$ eli yhteensä $3K$ ulostuloa.
- $K$:n arvo valitaan väliltä $\{0...K_{MAX}\}$ bayesilaisen informaatiokriteerin (BIC) avulla. $d$ on mallin parametrien määrä.

$$P(D|M_K) = P(D_l|\hat{\theta}, M_K) - \frac{d}{2}\log l$$

- Vaihtoehtoisesti voidaan vain testata erillisellä testidatalla millä $K$:n arvolla saadaan suurin uskottavuus
- Käytännössä mallin kompleksisuuden valinta on niin raskasta, että se tehtiin vain kerran käyttäen kaikkia näytteitä

## Algoritmit

- Testattiin kahta eri algoritmia sekä oikeilla että tarkoitusta varten luoduilla datoilla.
- Algoritmi 1 : Diskretoidaan jatkuvat datat ja käytetään Cooper-Herskovitsin pisteytystä rakenteen valinnassa. Parametrien arvot estimoidaan käyttäen neuroverkkoja ja alkuperäistä dataa.
- Algoritmi 2 : Rakenteen valinta suoritetaan neuroverkkopohjaisella pisteytyksellä ja parametrit estimoidaan samalla tavalla kuin ensimmäisessä algoritmissa.

# Algoritmien vertailu - 1

- Pyrittiin löytämään annetun luokkamuuttujan vanhemmat. Mitattiin sekä ennustustarkkuutta että kykyä löytää oikeat riippuvuussuhteet.

- Tarkkuus arvioitiin käyttämällä todellisia datasettejä (AUTO-MPG ja ABALONE). Mittareina käytettiin neliöllistä virhettä ja ehdollisen todennäköisyyden uskottavuutta

$$
\begin{aligned}
MSE &= \frac{1}{L} \sum_{D^{test}} (y^{(l)} - \hat{y}(\pi_y^{(l)}))^2 \\
LS &= -logP(D^{test}) = -\sum_{D^{test}} logP(y^{(l)} | \pi_y^{(l)})
\end{aligned}
$$

- $\hat{y}(\pi_y^{(l)})$ on posterioritodennäköisyyden odotusarvo

## Algoritmien vertailu - 2

- Oikean rakenteen valitsemista tutkittiin luomalla satunnaisia bayesverkkoja ja generoimalla näytteet niistä. Algoritmeilla saatuja verkkorakenteita verrattiin todelliseen malliin.
- Kahdeksan muuttujaa joista yksi valittiin luokkamuuttujaksi, $n$ muuttujaa sen vanhemmiksi ja loput näiden vanhemmiksi. Eri kokeissa $n$:n arvoa vaihdeltiin välillä 1..4.
- Ehdolliset todennäköisyydet generoitiin gaussisista sekoitemalleista joiden komponenttien määrää varioitiin välillä 3..7. Jakaumien parametrit riippuivat lineaarisesti vanhempien arvoista.
- Algoritmissa 1 diskretointi suoritettiin joko kahteen tai kolmeen arvoon. Algoritmissa 2 komponenttien maksimimäärä $K_{MAX}$ oli välillä 1..5 ja sen arvo valittiin BIC:llä.

## Tulokset

- Verkkorakennetta tutkittiin laskemalla montako kaarta algoritmit lisäsivät ja poistivat todelliseen malliin verrattuna.
- Saadut tulokset analysoitiin tilastollisilla testeillä.
- Ennustustarkkuuden erot eri algoritmien välillä eivät olleet tilastollisesti merkitseviä kummallakaan mittarilla. Algoritmilla 2 varianssi oli suurempi.
- Sen sijaan rakenteelliset erot olivat merkitseviä. Algoritmi 2 valitsee suuremman määrän vanhempia kuin algoritmi 1.
- Kaarien määrän vaihtelu oli suurempaa algoritmilla 2.

## Yhteenveto

- Neuroverkkoon pohjautuva pisteytysmenetelmä ei välttämättä ole kovin stabiili vaan saattaa juuttua paikallisiin maksimeihin. Tämä selittäisi tulosten suuren varianssin.
- Koska algoritmi 1 on lisäksi huomattavasti nopeampi on sen käyttö suositeltavaa. Algoritmin 2 suoritus kesti noin $30 - 40$ kertaisen ajan algoritmiin 1 verrattuna.
- Molemmat algoritmit ovat kohtuullisen raskaita koska jakaumat estimoidaan joka tapauksessa neuroverkoilla.