Usuzování v Bayesovských sítích pomocí multilineárních polynomů

P. Savický (nové výsledky společně J. Vomlel)

ÚI AV ČR, ÚTIA AV ČR

Seminář umělé inteligence, MFF UK, 29. května 2008

Holmesův a Watsonův trávník

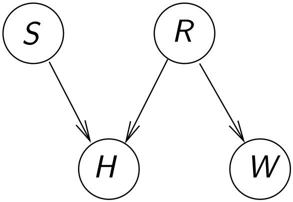
P. Holmes má na svém trávníku postřikovač, který někdy na noc zapíná.
Pokud je postřikovač zapnutý, tak je většinou trávník ráno mokrý.
P. Watson nemá na svém trávníku postřikovač, ale soused ano a někdy jeho trávník přes noc pokropí.
Když v noci pršelo, jsou ráno oba trávníky vždy mokré.

Když je ráno Holmesův trávník mokrý, jaká je pravděpodobnost, že byl Holmesův postřikovač zapnut?

Jak se tato pravděpodobnost změní, když navíc zjistíme, že je mokrý i Watsonův trávník?

Popis pomocí závislostí náhodných veličin

Zavedeme náhodné veličiny s hodnotami ano/ne:
$S \Leftrightarrow$ Holmesův postřikovač (sprikler) byl zapnutý
$R \Leftrightarrow$ v noci pršelo
$H \Leftrightarrow$ Holmesův trávník mokrý
$W \Leftrightarrow$ Watsonův trávník mokrý



(Holmes nesleduje předpověď počasí nebo je tato nezávislá na skutečnosti.)

Značení

Hodnoty S jsou $\overline{s}$ a s, analogicky pro další proměnné.

$\mathrm{P}(S)$ znamená funkci, která každé hodnotě S přiřadí její pravděpodobnost, v našem případě $\mathrm{P}(\overline{s})$ a $\mathrm{P}(s)$.

$\mathrm{P}(R|S)$ znamená funkci, která každé hodnotě R a S přiřadí příslušnou pravděpodobnost, tedy hodnoty $\mathrm{P}(\overline{r}|\overline{s}), \mathrm{P}(r|\overline{s}), \mathrm{P}(\overline{r}|s), \mathrm{P}(r|s)$.

$\mathrm{P}(R|S) = \mathrm{P}(R)$ je identita, která platí pro každé dosazení konkrétních hodnot za r, s.

Tabulky (podmíněných) pravděpodobností

$$\mathrm{P}(s) = 0.1$$

$$\mathrm{P}(r) = 0.2$$

$$\mathrm{P}(h|SR) = \begin{array}{cc} & \begin{array}{cc} \overline{r} & r \end{array} \\ \begin{array}{c} \overline{s} \\ s \end{array} & \begin{pmatrix} 0.0 & 1.0 \\ 0.9 & 1.0 \end{pmatrix} \end{array}$$

$$\mathrm{P}(w|R) = \begin{array}{c} \begin{array}{cc} \overline{r} & r \end{array} \\ \begin{pmatrix} 0.2 & 1.0 \end{pmatrix} \end{array}$$

Pravděpodobnosti $\mathrm{P}(\overline{s}), \mathrm{P}(\overline{r}), \mathrm{P}(\overline{h}|SR), \mathrm{P}(\overline{w}|R)$ jsou doplňky odpovídajících pravděpodobností do 1.

Od jednotlivých tabulek k společnému rozdělení S, R, H, W

Za jakých předpokladů platí

$$\mathrm{P}(SRHW) = \mathrm{P}(S)\ \mathrm{P}(R)\ \mathrm{P}(H|SR)\ \mathrm{P}(W|R)\ ? \tag{1}$$

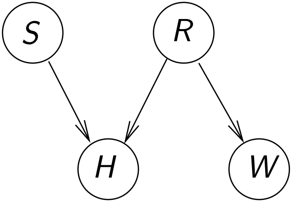
Součin podmíněných pravděpodobností vpravo definuje distribuci. Označme ji D. Pak platí

$$\begin{aligned} \mathrm{P}_D(S) &= \mathrm{P}(S) \\ \mathrm{P}_D(R) &= \mathrm{P}(R) \\ \mathrm{P}_D(H|SR) &= \mathrm{P}(H|SR) \\ \mathrm{P}_D(W|R) &= \mathrm{P}(W|R) \end{aligned} \tag{2}$$

D není jediná distribuce, která splňuje (2).

Platnost identity (1) je ekvivalentní (2) plus něco z následujícího
- některé podmíněné nezávislosti mezi zúčastněnými veličinami,
- předpoklad maximální entropie.

Příklady podmíněné nezávislosti



H, W jsou závislé, $\mathrm{P}(W|H) \neq \mathrm{P}(W)$.
Platí ale $\mathrm{P}(W|HR) = \mathrm{P}(W|R) \quad \Rightarrow$ podmíněná nezávislost.
Symetrický vztah, zapisuje se $H \perp W|R$.

K odvození souhrnné pravděpodobnosti budeme potřebovat:

$$\begin{array}{lll} S \perp R & \text{tj.} & \mathrm{P}(R|S) = \mathrm{P}(R) \\ SH \perp W|R & \text{tj.} & \mathrm{P}(W|SHR) = \mathrm{P}(W|R) \end{array}$$

Použití řetězcového pravidla

Pro libovolné veličiny (v libovolném uspořádání)

$$\begin{aligned}\mathrm{P}(SRHW) = \; & \mathrm{P}(S)\cdot \\ & \mathrm{P}(R|S)\cdot \\ & \mathrm{P}(H|SR)\cdot \\ & \mathrm{P}(W|SRH)\end{aligned}$$

S využitím podmíněných nezávislostí z předchozího slidu

$$\begin{aligned}\mathrm{P}(SRHW) = \; & \mathrm{P}(S)\cdot \\ & \mathrm{P}(R)\cdot \\ & \mathrm{P}(H|SR)\cdot \\ & \mathrm{P}(W|R)\end{aligned}$$

Toto je vyjádření pomocí podmíněných pravděpodobností, jejichž znalost předpokládáme.

Byl Holmesův postřikovač zapnutý?

S	R	H	W	pravděpodobnost
$\overline{s}$	$\overline{r}$	$\overline{h}$	$\overline{w}$	$0.9 \cdot 0.8 \cdot 1.0 \cdot 0.8$
$\overline{s}$	$\overline{r}$	$\overline{h}$	w	$0.9 \cdot 0.8 \cdot 1.0 \cdot 0.2$
$\overline{s}$	r	h	w	$0.9 \cdot 0.2 \cdot 1.0 \cdot 1.0$
s	$\overline{r}$	$\overline{h}$	$\overline{w}$	$0.1 \cdot 0.8 \cdot 0.1 \cdot 0.8$
s	$\overline{r}$	$\overline{h}$	w	$0.1 \cdot 0.8 \cdot 0.1 \cdot 0.2$
s	$\overline{r}$	h	$\overline{w}$	$0.1 \cdot 0.8 \cdot 0.9 \cdot 0.8$
s	$\overline{r}$	h	w	$0.1 \cdot 0.8 \cdot 0.9 \cdot 0.2$
s	r	h	w	$0.1 \cdot 0.2 \cdot 1.0 \cdot 1.0$

$$\mathrm{P}(s) = 0.1$$

$$\mathrm{P}(s|h) = \frac{\mathrm{P}(sh)}{\mathrm{P}(h)} = 0.3382$$

$$\mathrm{P}(s|hw) = \frac{\mathrm{P}(shw)}{\mathrm{P}(hw)} = 0.1604$$

Obecná Bayesovská síť

Předpokládáme náhodné veličiny $X_1, \ldots, X_n$,
které tvoří vrcholy orientovaného acyklického grafu.

$N = \{1, \ldots, n\}$.

$pa(i)$ je množina indexů bezprostředních předchůdců proměnné X_i.

Pro $A \subseteq N$ je X_A podmnožina proměnných sítě.

Dále předpokládáme, že je splněna identita

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathrm{P}(X_i | X_{pa(i)})$$

Usuzování v Bayesovské síti

Evidencí rozumíme znalost hodnot některých proměnných. Reprezentace přiřazením stavů x_A^* proměnným X_A, $A \subseteq N$.

Usuzováním rozumíme výpočet podmíněných pravděpodobností $\mathrm{P}(X_i = x_i^* | x_A^*)$ pro $i \notin A$.

Stačí odvodit postup výpočtu pravděpodobnosti libovolné evidence, protože platí

$$\mathrm{P}(X_i = x_i^* | x_A^*) = \frac{\mathrm{P}(x_i^*, x_A^*)}{\mathrm{P}(x_A^*)}$$

Klasická technika – konstrukce junction tree, message passing algoritmus.

Alternativa (Darwiche 2002) – použití polynomů reprezentovaných aritmetickým obvodem, jehož velikost je minimalizována určitou heuristikou.

Indikátorové proměnné

Nechť $X_1, \ldots, X_n$ jsou diskrétní náhodné proměnné.
Obor hodnot proměnné X_i označme V_i a $v_i = |V_i|$.

Proměnné X_i přiřadíme skupinu v_i indikátorových proměnných λ_x, jednu pro každé $x \in V_i$ (obvyklé značení je λ_{x_i} pro $x_i \in V_i$).

Přesněji (ale méně pohodlně):
Pokud $V_i = \{x_{i,1}, \ldots, x_{i,v_i}\}$, pak X_i odpovídá skupina proměnných

$$\lambda_{V_i} = \{\lambda_{x_{i,1}}, \ldots, \lambda_{x_{i,v_i}}\}$$

Stav x^* proměnné X_i budeme kódovat tak, že pro $x_i \in V_i$

$$\lambda_{x_i} = \begin{cases} 1 & \text{pokud} \quad x_i = x^* \\ 0 & \text{pokud} \quad x_i \neq x^* \end{cases}$$

Polynom odpovídající Bayesovské síti

Nechť $X_1, \ldots, X_n$ tvoří vrcholy orientovaného acyklického grafu a splňují podmíněné nezávislosti určené tímto grafem, tj.

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathrm{P}(X_i | X_{pa(i)})$$

Uvážíme polynom (sčítá se přes $V_1 \times \ldots \times V_n$) od indikátorových proměnných

$$f(\lambda_{V_1}, \ldots, \lambda_{V_n}) = \sum_{x_1, \ldots, x_n} \lambda_{x_1} \ldots \lambda_{x_n} \prod_{i=1}^{n} \mathrm{P}(x_i | x_{pa(i)})$$

Pokud je v každém bloku λ_{V_i} právě jedna jednička, je hodnotou polynomu pravděpodobnost odpovídající kombinace hodnot všech proměnných sítě,

Pokud jsou všechny indikátorové proměnné rovny 1, je polynom roven 1 (součet pravděpodobností všech kombinací).

Kódování obecné evidence

Pro libovolnou evidenci x_A^* pro $A \subseteq N$ (tj. událost $X_A = x_A^*$) platí

$$\mathrm{P}(x_A^*) = f(\lambda_{V_1}, \dots, \lambda_{V_n}) = \sum_{x_1,\dots,x_n} \lambda_{x_1} \dots \lambda_{x_n} \prod_{i=1}^{n} \mathrm{P}(x_i | x_{pa(i)})$$

kde ohodnocení proměnných λ_{V_i} je pro každé i definováno následovně:

Pokud $i \in A$, pak $\lambda_{x_i^*} = 1$ a ostatní v λ_{V_i} jsou rovny 0.
(λ_{V_i} kóduje $X_i = x_i^*$)

Pokud $i \notin A$, pak všechny proměnné λ_{V_i} jsou rovny 1.
(λ_{V_i} kóduje neznalost X_i)

Lze počítat f efektivně?

Obecně ne, ale v některých důležitých případech ano.

V těchto případech reprezentuje aritmetický obvod pro f exponenciální počet monomů díky střídání hladin sčítání a násobení.

Darwiche 2003: Pokud známe junction tree (získané nějakou standardní technikou), pak lze pro f najít aritmetický obvod složitosti lineární ve velikosti tabulek v tomto junction tree, tedy srovnatelné se složitostí message passing algoritmu.

Darwiche 2002: Aritmetický obvod může reprezentovat výhodnější výpočet, pokud existuje, např. díky tomu, že velké tabulky v síti mají pravidelnou strukturu.

Pro hledání obvodů navrhl Darwiche efektivní heuristickou metodu založenou na reprezentaci polynomu f pomocí Booleovské funkce.

Konstrukce obvodu pomocí Booleovského modelu

Zavedeme proměnné $\theta_{x_i|x_{pa(i)}}$ pro podmíněné pravděpodobnosti

$$\begin{aligned} f(\lambda_{V_1}, \ldots, \lambda_{V_n}) &= \sum_{x_1,\ldots,x_n} \lambda_{x_1} \ldots \lambda_{x_n} \prod_{i=1}^{n} \mathrm{P}(x_i|x_{pa(i)}) \\ &= \sum_{x_1,\ldots,x_n} \lambda_{x_1} \ldots \lambda_{x_n} \prod_{i=1}^{n} \theta_{x_i|x_{pa(i)}} \end{aligned}$$

Koeficienty polynomu jsou $0, 1$ a polynom je multilineární.
Každý monom je určen množinou proměnných.
Polynom je množina některých monomů.

Reprezentace polynomu pomocí Booleovské funkce

Každé proměnné λ_{x_i} a $\theta_{x_i|x_{pa(i)}}$ přiřadíme Booleovskou proměnnou $\lambda^B_{x_i}$ a $\theta^B_{x_i|x_{pa(i)}}$.

Polynom

$$f(\lambda_{V_1}, \ldots, \lambda_{V_n}) = \sum_{x_1,\ldots,x_n} \lambda_{x_1} \ldots \lambda_{x_n} \prod_{i=1}^{n} \theta_{x_i|x_{pa(i)}}$$

pak reprezentujeme Booleovskou funkcí g definovanou následovně.

$g($ všechny $\lambda^B_{x_i}$, všechny $\theta^B_{x_i|x_{pa(i)}}) \Leftrightarrow$ monom určený ohodnocením vstupních proměnných je v polynomu f.

Výchozí popis g je pomocí CNF.

Konstrukce obvodu pro f pomocí g

Darwiche 2002: Jestliže pro g najdeme reprezentaci v tzv. d-DNNF (deterministic decomposable negation normal form) pak lze pro f najít obvod stejné velikosti.

Pro tuto formu lze SAT testovat v lineárním čase $\Rightarrow$ převod z CNF obecně vyžaduje exponenciální nárůst velikosti, ale ne vždy.

Pro hledání d-DNNF navrhl Darwiche heuristiku vycházející z heuristik pro test splnitelnosti (Chaff).

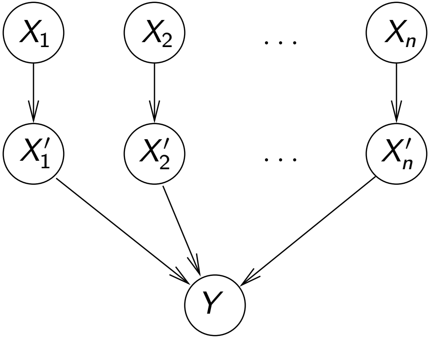
Je k dispozici program Ace autorů Darviche, Chavira, který implementuje uvedenou heuristiku v Java.

Reprezentace noisy-OR s pomocnými uzly

Vstupy $X_1, \ldots, X_n$, výstup $Y = X_1' \vee \ldots \vee X_n'$, kde $X_i' \leq X_i$ jsou kopie X_i ovlivněné šumem ($p_i < 1$):

$$\mathrm{P}(X_i' = 0 | X_i = 0) = 1 = p_i^0$$

$$\mathrm{P}(X_i' = 0 | X_i = 1) = p_i = p_i^1$$



Tabulky parametrů pro noisy-OR

Celková závislost je vyjádřena formulí

$$\mathrm{P}(Y = 0|x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i^{x_i}$$

Jde o zobecnění deterministické funkce OR (uvažujeme $0^0 = 1$):

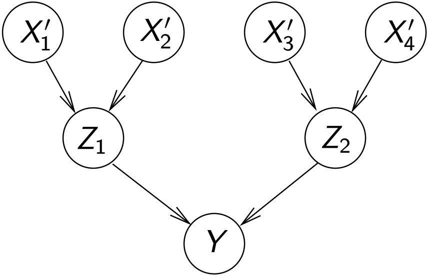
$$\mathrm{P}(Y = 0|x'_1, \ldots, x'_n) = \prod_{i=1}^{n}(1 - x'_i) = \prod_{i=1}^{n} 0^{x'_i}$$

Pomocné uzly lze uvažovat explicitně nebo je vyeliminovat.
V obou případech máme tabulku s exponenciálně mnoho členy.

Rozklad pomocí binárního stromu

Standarní transformace (považuji za folklore).

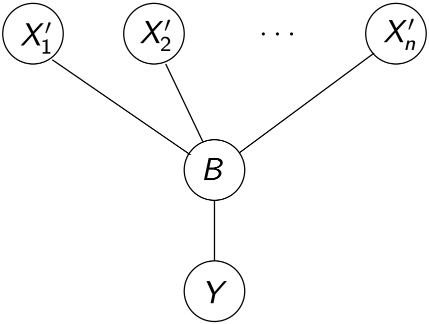
Pro jednoduchost uveďme verzi s pomocnými uzly a pro $n = 4$.



Polynomiální počet parametrů v souhrnu tabulek.

Rozklad pomocí neorientovaného grafu

Díez, Galán, 2003 navrhli rozklad noisy-max (tedy i noisy-or) s pomocnou proměnnou B



$$\mathrm{P}(Y = y | X_1 = x'_1, \dots, X_n = x'_n) = \sum_{b=0}^{1} \xi(b, y) \cdot \prod_{i=1}^{m} \varphi_i(b, x'_i)$$

Stanovení interakcí na hranách

Tabulky v rozkladu

$$P(Y = y|X_1 = x_1', \ldots, X_n = x_n') = \sum_{b=0}^{1} \xi(b, y) \cdot \prod_{i=1}^{m} \varphi_i(b, x_i')$$

je třeba zvolit tak, aby výše uvedená identita byla totožná s identitou (pravé strany obou jsou součet dvou součinů)

$$P(Y = y|X_1 = x_1', \ldots, X_n = x_n') = (1 - 2y) \prod_{i=1}^{n} (1 - x_i') + y \prod_{i=1}^{n} 1$$

Rozklad na tenzory ranku 1

Rozklad

$$P(Y = y | X_1 = x'_1, \dots, X_n = x'_n) = \sum_{b=0}^{1} \xi(b, y) \cdot \prod_{i=1}^{m} \varphi_i(b, x'_i)$$

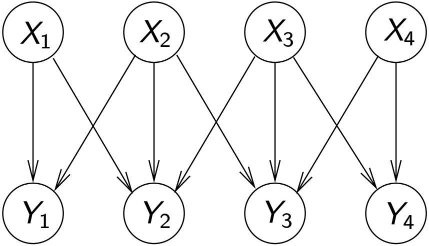
je rozklad na tenzory ranku 1 (Savický, Vomlel, 2006)

Pro noisy-or je rank 2, proto B má dvě hodnoty.
Optimalita i pro noisy-max.
Existují explicitní rozklady i pro některé jiné funkce
Obecně velmi obtížné hledat

Srovnání binárních stromů a tenzorových rozkladů

Pro jednotlivé noisy-or uzly není rozdíl příliš výrazný.

Pro sítě typu BN2O (Y_i jsou noisy-or) bez předpokladu známé evidence může být rozdíl citelný (Vomlel, Savický, 2008).



Používají se např. v lékařské diagnostice.
Při známé evidenci - algoritmus QuickScore.

Popis experimentu

Parametry BN2O sítí:
x počet vstupních uzlů, např. $10, 20, 30$.
y počet výstupních uzlů, např. $10, 15, 20, 30, 40$.
e počet hran, např. $60, 100, 200$.
Pro vybrané kombinace x–y–e jsme vygenerovali náhodně síť.

Program Ace (Darwiche, Chavira) pro kompilaci sítě do obvodu. (32 bit, Java)

Kompilace sítí do obvodu dvěma způsoby:

- Kompilace originální sítě (Ace používá rozklad pomocí binárních stromů).
- Kompilace transformované sítě s využitím tenzorového rozkladu.

Transformace je úspěšná $\Leftrightarrow$ vede na menší výsledný obvod.

Výsledky experimentu

x–y–e	e/x	originální	transformovaná	o/t
10–1–10	1.00	268	182	1.47
50–20–80	1.66	6 487	3 843	1.69
10–15–60	6.00	7 167	5 261	1.36
6–40–128d	21.33	8 173	6 054	1.35
30–10–100	3.33	1 050 339	10 608	99.0
10–30–100d	10.00	13 085	10 899	1.20
10–15–90	9.00	27 096	38 612	0.70
30–15–150	5.00	230 155 117	298 612	771.
20–30–150	7.50	899 979	890 743	1.01
20–30–151d	7.55	1 457 044	1 272 300	1.15
20–40–196d	9.80	1 993 531	1 970 983	1.01
50–20–200	4.00	N/A	3 075 737	
30–20–200	6.67	N/A	8 232 072	
20–30–300	15.00	21 110 387	53 692 211	0.39
50–50–200	4.00	73 734 475	68 006 871	1.08
100–100–400	4.00	N/A	N/A	