Efficient Relational Learning with Hidden Variable Detection

Ni Lao, Jun Zhu, Liu Liu, Yandong Liu, William W. Cohen

Statistical Relational Learning

Relational tasks are everywhere

collective classification (Tasker et al. 2002)

information extraction (Poon & Domingos 2007; Bunescu 2004)

social network modeling (Kemp et al. 2006)

Modeling Long Rang Dependency (LRD) is hard ☹

e.g. smokes(A)& friends (A,B))& friends (B,C) ➔ smokes(C)

e.g. IsMotherOf(A,B) ➔¬ IsFatherOf(A,C)

The Markov blanket of a variable grows prohibitively fast as the model's order of Markov dependency grows.

Discovering hidden roles help capture LRD ☺

e.g. topic models, block models

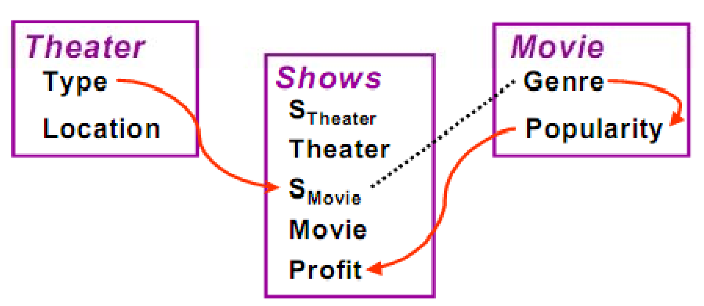
Thus reduce the need of extensive structure learning

Bayesian Networks vs. MRFs

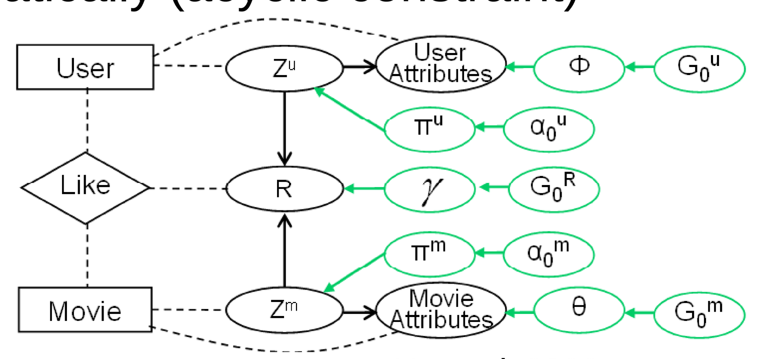
Relational Bayesian Networks

☺ Easy to do inference and learning

☹ Cannot learn structure automatically (acyclic constraint)



Getoor et al. 2001

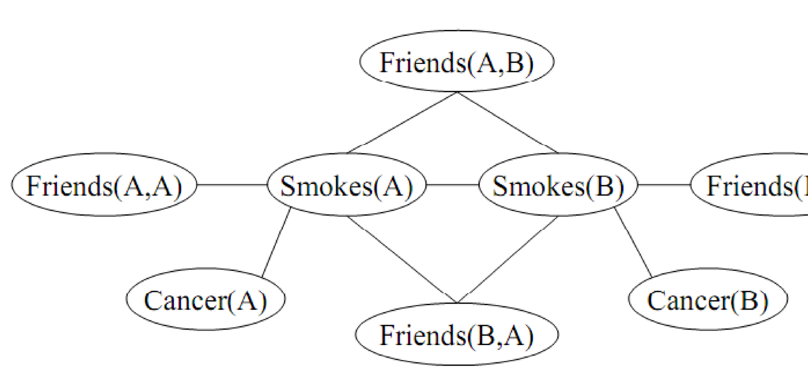


Xu et al. 2006

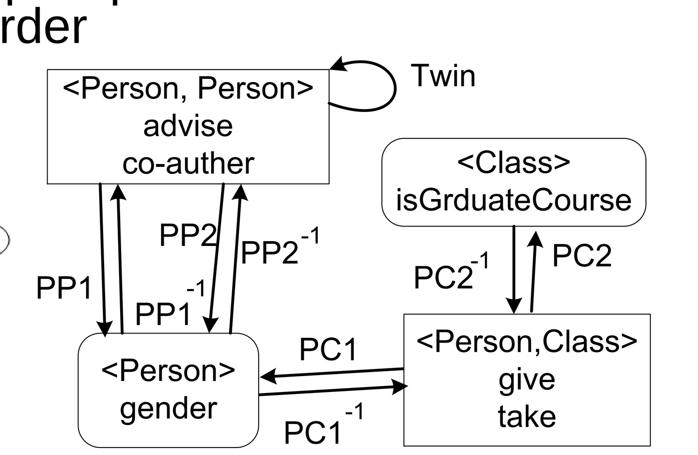
Relational Markov Networks (RMNs)

☺ Flexibility in representing complex patterns

☹ Inference and learning are harder



Kok & Domingos 2005



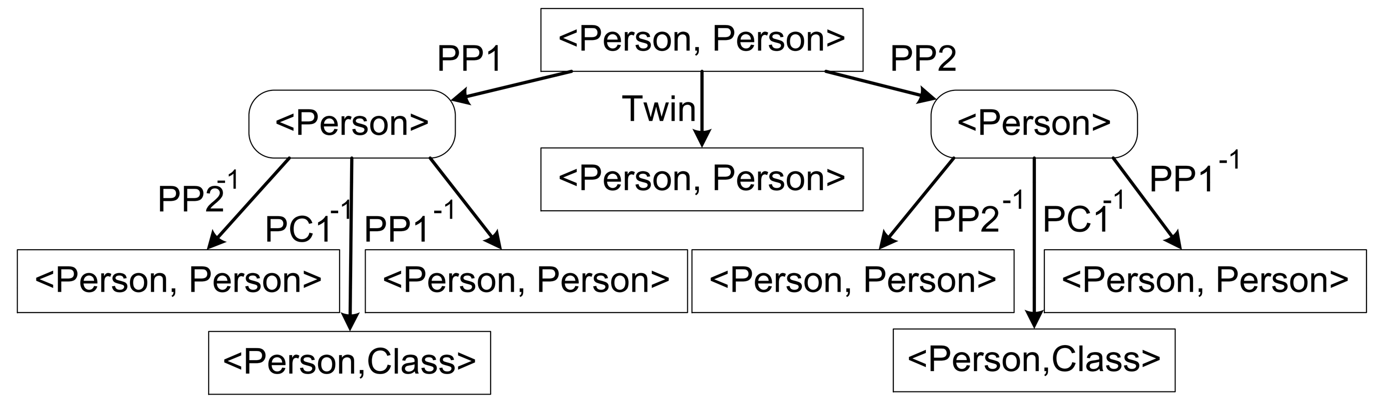
A schema for the university domain (this work)

For efficiency, we only consider pair wise features, and we use mean field contrastive divergence (Welling & Hinton, 2001) to do approximately optimize a regularized objective

$$L(\theta) = \log \sum_{\mathbf{h}} p(\mathbf{y},\mathbf{h} \mid \mathbf{o};\theta) - \lambda |\theta|_1 - \beta |\theta|_2^2$$

Tree RMN

The Markov blanket of an entity *e* can be concisely defined by a relation tree starting from its type node



Two-level relation trees for the <Person, Person> type

Variable Induction

Adding variables as needed

Algorithm 1 Contrastive Variable Induction

initialize a treeRMN $\mathcal{M} = (G, \mathbf{f}, \theta)$
while true **do**
  estimate parameters θ by L-BFGS
  $(\mathbf{f}', \theta') = \mathit{induceHiddenVariables}(\mathcal{M})$
  if no hidden variable is induced **then**
    break
  end if
end while
return $\mathcal{M}$

How to *efficiently* evaluate a candidate *H*?

2nd order Taylor expansion estimates that each new feature *f*

bring maximum gain $\Delta_{I,f} = \frac{1}{2}\frac{\lfloor -e_I[f] \rfloor_\lambda^2}{\delta_I[f]+\beta}$ at $\theta_f = \frac{\lfloor -e_I[f] \rfloor_\lambda}{\delta_I[f]+\beta}$

where *I* is the set of entities with *H=1*. The overall gain is

$$\Delta_I \approx \sum_{f \in \mathbf{f}_I} \Delta_{I,f}$$

How to *efficiently* sift through all candidates?

We use a naïve bottom up clustering algorithm

Algorithm 2 Bottom Up Clustering of Entities

initialize clustering $\Gamma = \{I_i = \{i\}\}$
while true **do**
  for any pair of clusters $I_1, I_2 \in \Gamma$ **do**
    $inc(I_1, I_2) = \Delta_{I_1 \cup I_2} - \Delta_{I_1} - \Delta_{I_2}$
  end for
  if the largest increment ≤ 0 **then**
    break
  end if
  merge the pair with the largest increment
end while
return Γ

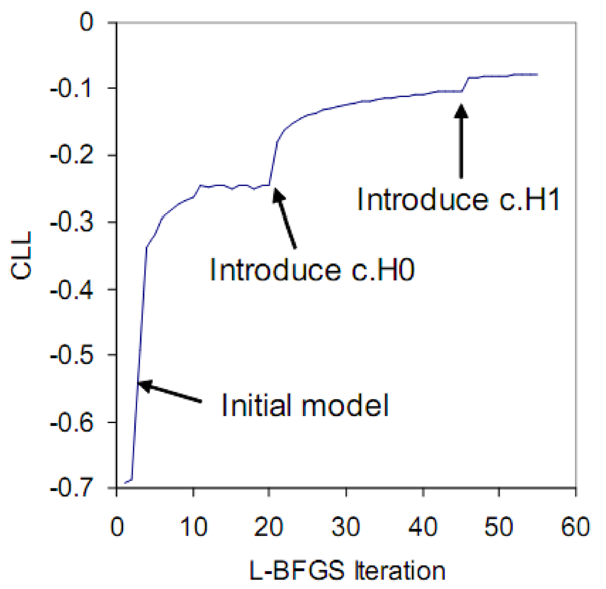
Results

The datasets and a learning curve

	Basic #E	Basic #A	Composite #E	Composite #A
Animal	50	80	0	0
Nation	14	111	196	56
UML	135	0	18,225	49
Kinship	104	0	10,816	1*

Table 1: Number of entities (#E) and attributes (#A) for four datasets. *The kinship data has only one attribute which has 26 possible values.

Previous approaches
MLN structure learning (MLS) [10],
Infinite Relational Models (IRM) [9]
Multiple Relational Clustering (MRC) [11]



Example hidden variables

Animal data

	Entities	Positive Features	Negative Features
C0	KillerWhale Seal Dolphin BlueWhale Walrus HumpbackWhale	Flippers Ocean Water Swims Fish Hairless Coastal Arctic ...	Quadrapedal Ground Furry Strainteeth Walks ...
C1	GrizzlyBear Tiger GermanShepherd Leopard Wolf Weasel Raccoon Fox Bobcat Lion	Stalker Fierce Meat Meatteeth Claws Hunter Nocturnal Paws Smart Pads ...	Timid Vegetation Weak Grazer Toughskin Hooves Domestic ...
C2	Hamster Skunk Mole Rabbit Rat Raccoon Mouse	Hibernate Buckteeth Weak Small Fields Nestspot Paws ...	Strong Muscle Big Toughskin ...
C3	SpiderMonkey Gorilla Chimpanzee	Tree Jungle Bipedal Hands Vegetation Forest ...	Plains Fields Patches ...

UML data

	Entities	Positive Features
C0	AcquiredAbnormality AnatomicalAbnormality CongenitalAbnormality	c $\xrightarrow{CC2^{-1}}$ cc.Causes c $\xrightarrow{CC1^{-1}}$ cc.PartOf c $\xrightarrow{CC2^{-1}}$ cc.Complicates c $\xrightarrow{CC2^{-1}}$ cc.CooccursWith ...
C1	Alga Plant	c $\xrightarrow{CC1^{-1}}$ cc.InteractsWith c $\xrightarrow{CC1^{-1}}$ cc.LocationOf ...
C2	Amphibian Animal Bird Invertebrate Fish Mammal Reptile Vertebrate	c $\xrightarrow{CC1^{-1}}$ cc.InteractsWith c $\xrightarrow{CC2^{-1}}$ cc.PropertyOf c $\xrightarrow{CC2^{-1}}$ cc.InteractsWith c $\xrightarrow{CC2^{-1}}$ cc.PartOf ...

Main result

Animal, λ=0.01, β=1					**Nation**, λ=0.01, β=1				
	CLL	**AUC**	**dim$_\theta$**	**Time**		**CLL**	**AUC**	**dim$_\theta$**	**Time**
RMN$_0$	-0.34±0.03	0.88±0.02	3,655	5s	**RMN$_0$**	-0.40±0.01	0.63±0.04	7,812	15s
					RMN$_1$	-0.33±0.02	0.72±0.04	21,840	70s
					RMN$_2$	-0.38±0.03	0.71±0.04	40,489	446s
RMN$_0^{CVI\star}$	**-0.33±0.02**	**0.89±0.02**	4,349	9s	**RMN$_1^{CVI}$**	**-0.31±0.02**	**0.83±0.04**	22,191	104s
MSL	-0.54±0.04	0.68±0.04		†24h	**MSL**	-0.33±0.04	0.77±0.04		†24h
MRC	-0.43±0.04	0.80±0.04		†10h	**MRC**	**-0.31±0.02**	0.75±0.03		†10h
IRM	-0.43±0.06	0.79±0.08		†10h	**IRM**	-0.32±0.02	0.75±0.03		†10h
UML, λ=0.01, β=10					**Kinship**, λ=0.01, β=10				
	CLL	**AUC**	**dim$_\theta$**	**Time**		**CLL**	**AUC**	**dim$_\theta$**	**Time**
RMN$_0$	-0.056±0.005	0.70±0.02	1,081	0.3h	**RMN$_0$**	§-2.95±0.01	0.08±0.00	25	6s
RMN$_1$	-0.044±0.002	0.68±0.04	2,162	1.0h	**RMN$_1$**	§-1.36±0.05	0.66±0.03	350	107s
RMN$_2$	-0.028±0.003	0.71±0.02	6,440	14.5h	**RMN$_2$**	§-2.34±0.01	0.33±0.00	1,625	2.1h
RMN$_1^{CVI\star}$	-0.005±0.001	0.94±0.01	6,946	453s	**RMN$_1^{CVI}$**	§-1.04±0.03	0.81±0.01	900	402s
MSL	-0.025±0.002	0.47±0.06		†24h	**MSL**	-0.066±0.006	0.59±0.08		†24h
MRC	**-0.004±0.000**	**0.97±0.00**		†10h	**MRC**	**-0.048±0.002**	**0.84±0.01**		†10h
IRM	-0.011±0.001	0.79±0.01		†10h	**IRM**	-0.063±0.002	0.68±0.01		†10h

Carnegie Mellon

{nlao,junzhu,liuliu,yandongl,wcohen}@cs.cmu.edu

--- in NIPS 2010 ---