Comparatives in Relational-Statistical Modelling and Learning for Bioinformatics

Nicos Angelopoulos and Stephen Muggleton

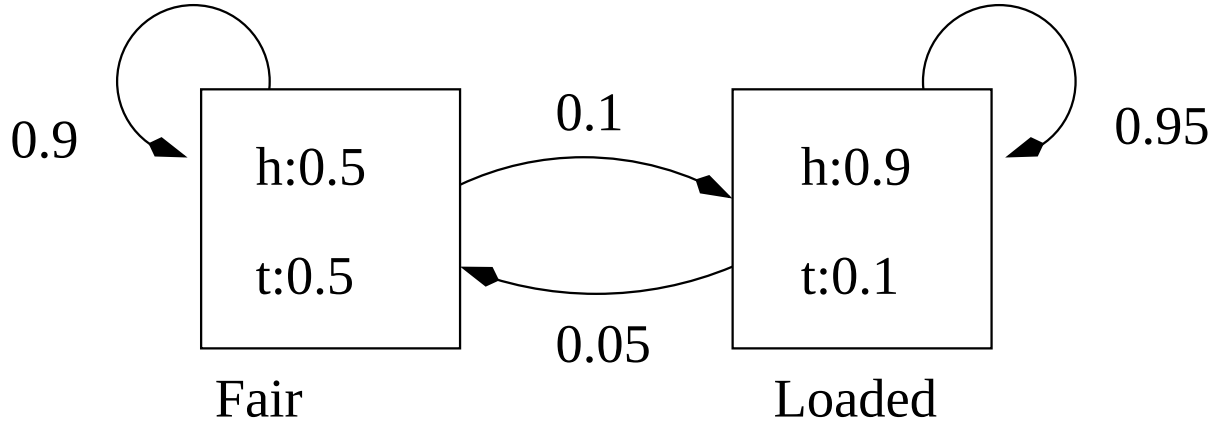
{nicos,shm}@doc.ic.ac.uk.

Imperial College, London.

talk structure

- HMMs
- Bayesian Networks
- PRMs
- Prism
- SLPs
- conclusions

HMMs



Hidden Markov Model example:

two coin states and two observations.

HMMs for ML

+ Representation of independent processes in physical systems. Efficient algorithms,
  - reasoning: Viterbi and forward/backward,
  - parameter learning: Baum-Welch.

- Restrictive independence condition. Not compositional ways of constructing states and transitions.

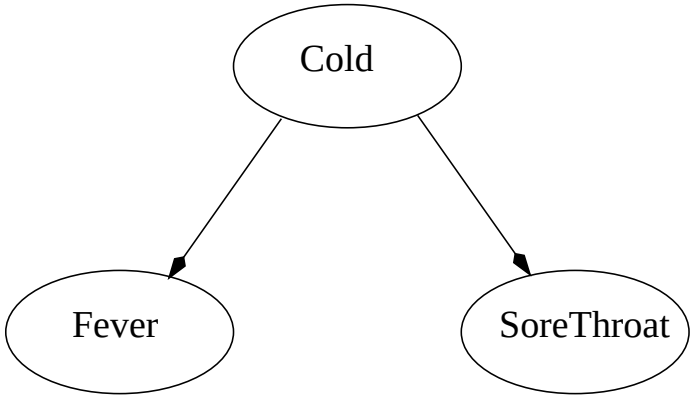
HMMs research

A. Krogh.
Gene finding: putting the parts together.
In, Guide to Human Genome Computing,Ch. 11, pp.261–274.

K. Asai, S. Hayamizu and K. Handa
Prediction of protein secondary structure by the hidden Markov model.
Comp. Appl. in the Biosciences, Vol.9, no.2, 141-146, 1993.

R. Durbin, S. Eddy, A. Krogh and G. Mitchison
Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1999.

BNs



		Cold	
		yes	no
Fever	yes	0.8	0.7
	no	0.2	0.3

		Cold	
		yes	no
SoreThroat	yes	0.9	0.8
	no	0.1	0.2

Bayesian Networks example:
simple cause effect of three two-valued variables.

BNs for ML

+ Clear probabilistic semantics. Well understood algorithms,
  - reasoning: probability updating, most probable explanation,
  - parameter and structure learning: batch learning, model selection.

- Propositional. Inflexibility of causality for logical modelling. Unable to reason with loops.

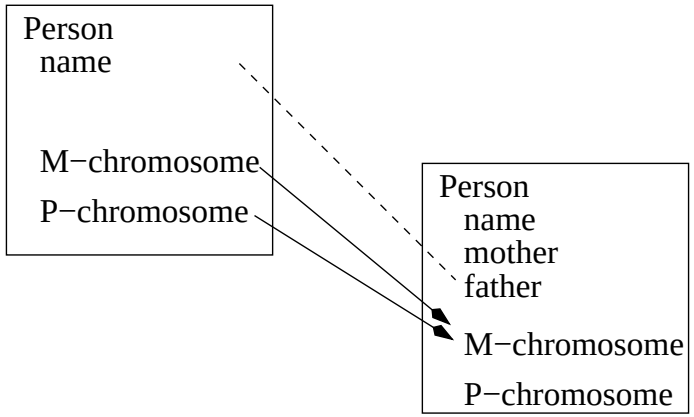
BNs research

D. Cai, B. Kao, S. Kasif, and A. Delcher.
Modeling splice sites with bayes networks.
Bioinformatics, 16(2):152–158, 2000.

S.C. Schmidler, J.S. Liu and D. L. Brutlag.
Bayesian protein secondary structure.
J. of Comp. Biology, 7(1/2):233–248, 2000.

N. Friedman, M. Linial, I. Nachman and D. Pe'er
Using Bayesian Networks to Analyze Expression Data.
Proc. of RECOMB 2000, 2000.

PRMs



Probabilistic relational model example,

probabilistic dependency of attributes.

PRMs for ML

+ Ability to work with available databases and structured but un-modelled data. Introduced algorithms for parameter estimation and some form of structural learning.

- Cannot add probabilities over processes. Not suitable as a probabilistic programming language.

PRMs research

N. Friedman and L. Getoor and D. Koller and A. Pfeffer

Learning Probabilistic Relational Models.

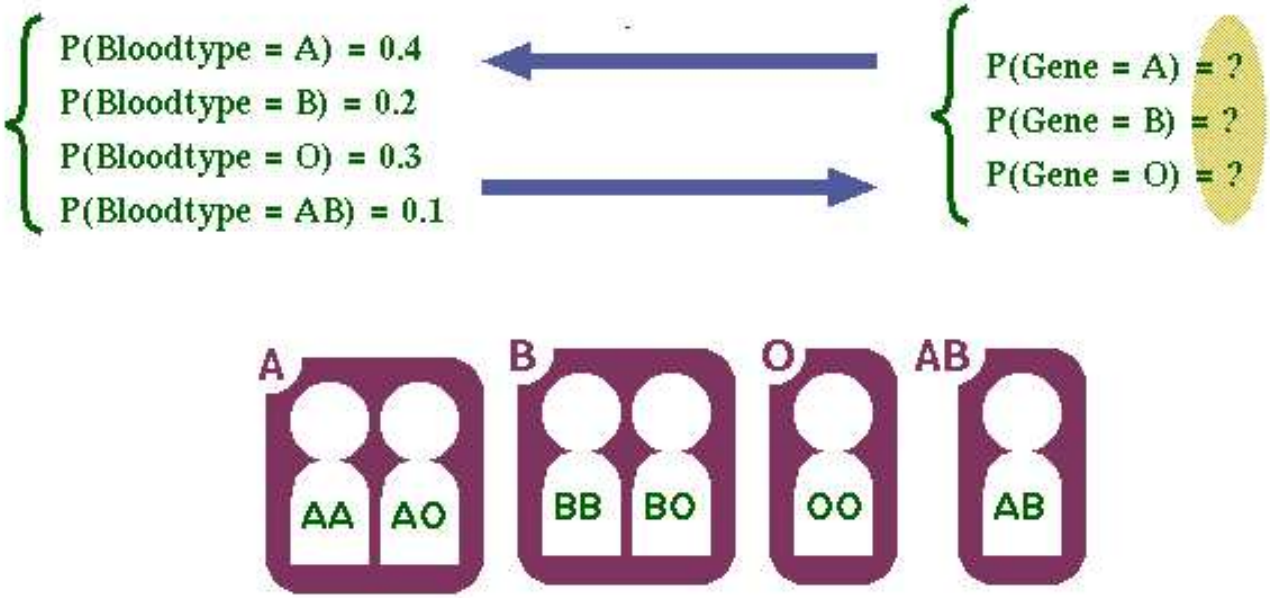
Proc. of 16th IJCAI, pp.1300-1307, 1999.

L. Getoor, D. Koller, B. Taskar, and N. Friedman

Learning Probabilistic Relational Models with Structural Uncertainty.

Proc. of AAAI2000 workshop on Learning Statistical Models from Relational Data, 2000.

Prism



Prism example,
from bloodtype to genes.

Prism cont.

(C1) bloodtype(a):- genotype(a,a); genotype(a,o); genotype(o,a).
(C2) bloodtype(b):- genotype(b,b); genotype(b,o); genotype(o,b).
(C3) bloodtype(o):- genotype(o,o).
(C4) bloodtype(ab):- genotype(a,b); genotype(b,a).
(C5) genotype(X,Y):- gene(father,X),gene(mother,Y).
(C6) gene(Parent,Gene):- msw(gn,Parent,Gene).
(C7) target(bloodtype,1).
(C8) values(gn,[a,b,o]).
(C9) data('bloodtype.dat').

Prism program,

from bloodtype to genes.

Prism for ML

+ Full expressive power of Logic Programming. Efficient EM algorithm for parameter estimation.

- Probability distributions are only allowed over proposition. No structure learning algorithm.

Prism research

T. Sato and Y. Kameya

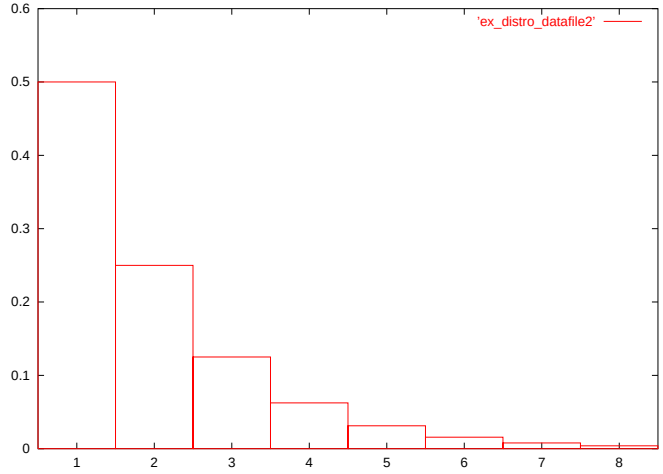
PRISM: A symbolic-statistical modeling language

Proc. of IJCAI97, pp.1330-1335, 1997.

T. Sato and Y. Kameya

Parameter Learning of Logic Programs for Symbolic-Statistical Modeling. *Journal of Artificial Intelligence Research*, pp.391-454, 2001.

SLPs




```
1/2 :: nat( 0 ).
1/2 :: nat( s(X) ) :- nat( X ).
```

Stochastic Logic Program example,

distribution of query on Program.

SLPs for ML

+ Full expressivity of Logic Programming. Probabilities over general if-then rules (clauses). Learning algorithms:
  - EM, for parameter estimation.
  - Some structure learning via MCMC.
- Probabilistic aspect not as compositional as logical part. No efficient most probable explanation, or answer, algorithm. Failure paths are very costly. No ILP-like structure learning.

SLPs research

N. Angelopoulos and S. Muggleton.
Machine learning metabolic pathway descriptions using a probabilistic relational representation.
In Machine Intelligence 19, September 2002.

J. Cussens
Parameter estimation in stochastic logic programs.
Machine Learning, 44(3):245–271, 2002.

N. Angelopoulos and J. Cussens.
Markov chain Monte Carlo using tree-based priors on model structure.
In UAI-2001, pp. 16-23, 2002.

conclusions

Simpler representation have been used to-date, due to their simplicity, the robustness and the efficiency of associated algorithms.

However, as the volume of data and the need to combine complex information increases, holistic representations will become more attractive.

Formalisms combining logic and probability have attractive features that may deem them strong candidates. The informatics challenge is to provide the algorithms that will power these richer representations.