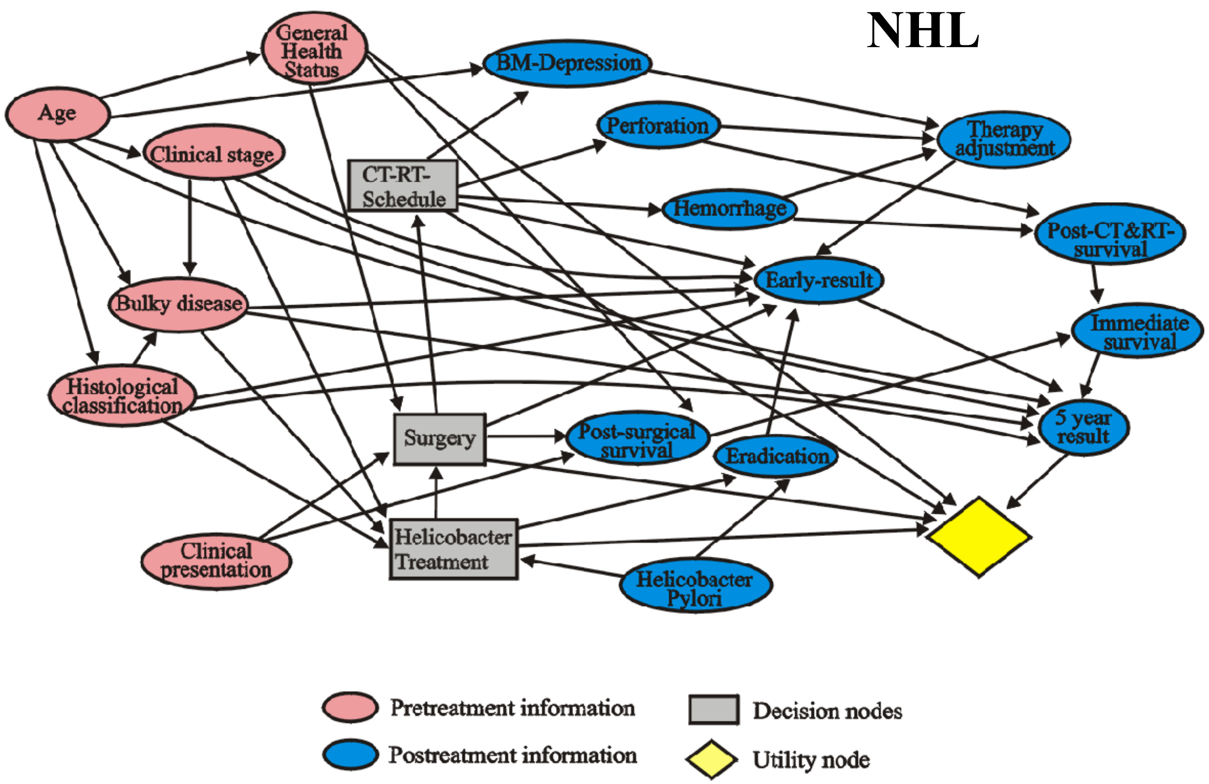
The useR! 2009 Conference in Rennes, July 8-10, 2009

Influence Diagrams on R



J.A. Fernández del Pozo; Contact author: jafernandez@fi.upm.es
C. Bielza
Departamento de Inteligencia Artificial,
Universidad Politécnica de Madrid, Spain

Influence Diagrams on R **IdR**

Keywords: Machine Learning,
Probabilistic Graphical Models,
Bayesian Networks,
Classification,
Decision Making

1. Decision making models:
   Bayesian Networks and Influence Diagrams
2. Decision model evaluations
3. Analysis and explanations of the results

An educational package for Probabilistic Graphical Models

http://www.dia.fi.upm.es/~jafernan/research/idr/idr.html
http://www.dia.fi.upm.es/~jafernan/research/idr/IdR_1.0.zip

YARP: Yet Another R Package?

Educational and Research purposes about Probabilistic Graphical Models using the R environment (RGui, lattice, cluster, gat,…)

•Grappa: R functions for probability propagation,
http://www.maths.bris.ac.uk/~peter/Grappa/
Peter J. Green
University of Bristol, UK

•Related R packages
CRAN Task View: gRaphical Models in R
deal, bnlearn, MASTINO,….

•Other Software (not in R):
GeNIe & SMILE: http://genie.sis.pitt.edu
Hugin: http://www.hugin.com
Elvira: http://www.leo.ugr.es/elvira/
Ace, the Bayesian network compiler: http://reasoning.cs.ucla.edu/ace
Mark Chavira and Adnan Darwiche
The Automated Reasoning group at UCLA

Decision making and classification problems

Influence diagrams (ID) and Bayesian network (BN) with discrete random variables as R scripts (pure R):

- Description of the model: graph + probability distributions
- Build the graph and assign the probability distributions
- Learn Bayesian networks from data and simulate data from the model
- Inference: optimal decision policies and marginal distributions
- Mining the results for validation and sensitivity analysis

•Decision support

Evaluation, simulation, learning, queries

1. Influence Diagram & Bayesian Network
   Bayes' theorem, Conditional expectation,
   Expected Utility maximization

2. Node definition
   Attributes
   Arcs and
   Probabilistic dependence
      Conditional probability tables:
         Probability distributions (matrix by rows)

```
node <- function( Node=NULL, ## copy
  Type=NULL, Name=NULL, Values=NULL, Preds=NULL, Pots=NULL, ## new
  Mpot=NULL, Maxszpot=0, EPSILON=1e-25,   trz.definition=FALSE)
```

Main functions

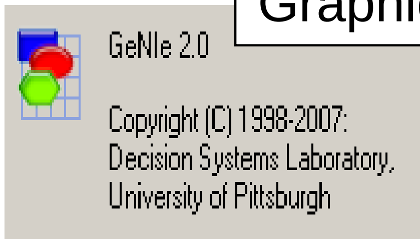
node: type, values, links, probabilistic dependences

influence.diagram (chance, decision and utility variables)

bayesian.network (chance variables)

…….

Graphical interface



Inference functions

remarc.eval and remnod.eval – ID exact evaluation

marnod.eval and sample.eval – BN exact and approximate inference

evid.inst – queries

…….

Core functions (graph and probability management)

check.rr: DAG conditions and other properties (c.c, mady)

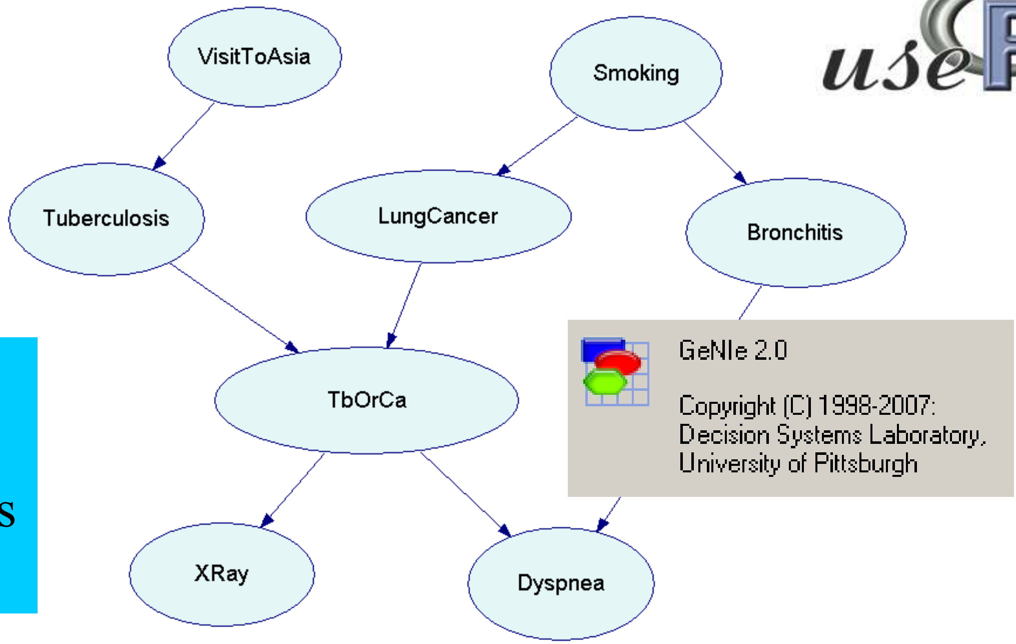
bayes.i, bayes.j – implementation of Bayes rule

conditional.expectation – combine utilities and uncertainty

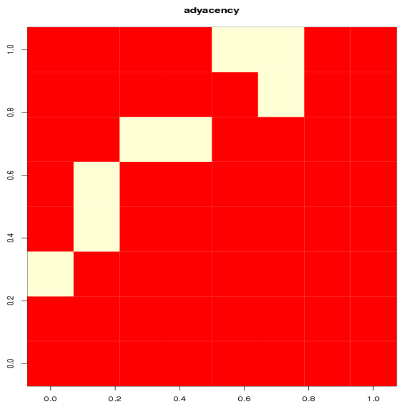
max.utility – define the optimal decision policy on every scenario

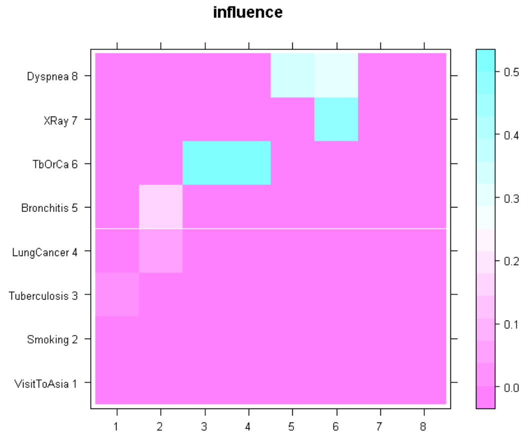
…….

Bayesian network **definition** and evaluation: Asia



BN represents the joint probability distribution over all variables using the chain rule and the independences expressed by the graph

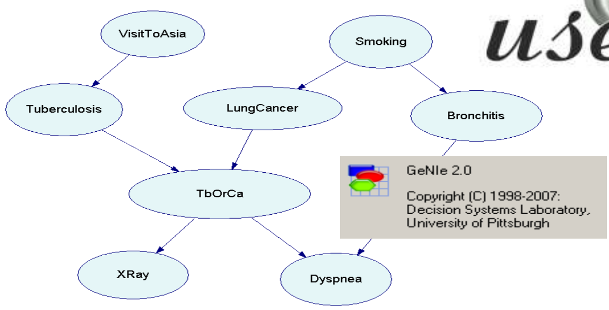





```
•Bronchitis = node( Type = "CHANCE",
Name = "Bronchitis",
Values=c("ABSENT","PRESENT"),
Preds=c("Smoking"),
Pots=matrix( data = c(
0.70, 0.30,
0.40, 0.60),
  nrow = 2, ncol = 2, byrow = TRUE,
 dimnames = NULL)),
```

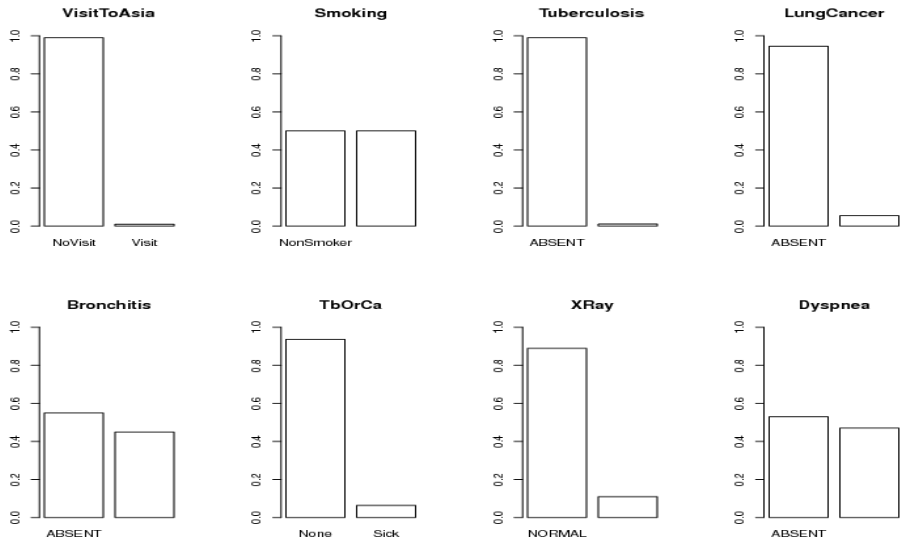
This is an example of graphical model useful in demonstrating basic concepts of Bayesian networks in diagnosis.
Lauritzen, Steffen L. & Spiegelhalter, David J. (1988).
Local computations with probabilities on graphical structures and their application to expert systems,
Journal of the Royal Statistical Society B, 50(2):157-224.

Bayesian network definition and **evaluation**: Asia



Approximate sampling

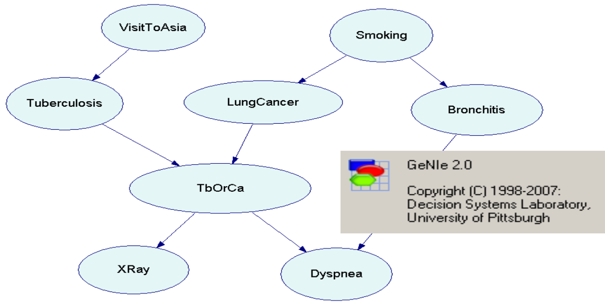
```
> B <- samnod.eval( asia)
Network summary: netname   Nodes:  8
Arcs:  8    Max preds:  2    Max succs:  2
Maxszpot:  18
TOTAL COMPLEX:  36   MAX COMPLEX:  8
Network summary: netname   Nodes:  8
Arcs:  8    Max preds:  2    Max succs:  2
Maxszpot:  18
TOTAL COMPLEX:  36   MAX COMPLEX:  8
SAMPLE EVALUATION:  netname   size:  36
mpd:  0.9897 0.0103    Node:  VisitToAsia
mpd:  0.4994 0.5006    Node:  Smoking
mpd:  0.9897 0.0103    Node:  Tuberculosis
mpd:  0.9478 0.0522    Node:  LungCancer
mpd:  0.5444 0.4556    Node:  Bronchitis
mpd:  0.9381 0.0619    Node:  TbOrCa
mpd:  0.1050 0.8950    Node:  XRay
mpd:  0.6050 0.3905    Node:  Dyspnea
```




```
[1] "CONDITIONALS."
# 1  ----- 0.99 0.01        ::--^--::   VisitToAsia
#  ------- NoVisit Visit    ::--|--::   Tuberculosis
# 2  ----- 0.50 0.50        ::--^--::   Smoking
#  ------- NonSmoker Smoker ::--|--::   LungCancer Bronchitis
# 3  ----- 0.989 0.0104     ::--^--::   Tuberculosis
#  ------- ABSENT PRESENT   ::--|--::   Bronchitis XRay Dyspnea TbOrCa
# 4  ----- 0.945 0.0550     ::--^--::   LungCancer
#  ------- ABSENT PRESENT   ::--|--::   Tuberculosis Bronchitis XRay Dyspnea TbOrCa
# 5  ----- 0.550 0.450      ::--^--::   Bronchitis
#  ------- ABSENT PRESENT   ::--|--::   XRay Dyspnea TbOrCa
# 6  ----- 0.946 0.054      ::--^--::   TbOrCa
#  ------- None Sick        ::--|--::
# 7  ----- 0.901 0.099      ::--^--::   XRay
#  ------- NORMAL ABNORMAL  ::--|--::   Dyspnea TbOrCa
# 8  ----- 0.528 0.471      ::--^--::   Dyspnea
#  ------- ABSENT PRESENT   ::--|--::   TbOrCa
```

Exact marginalization

Bayesian network definition and **evaluation**: Asia




```
> summary.network( asia, verbose=TRUE)
```

Network summary: netname   Nodes: 8
Arcs: 8   Max preds: 2   Max succs: 2

Node: **VisitToAsia**
  gr: 0   COMPLEX 2   Preds:
Node: **Smoking**
  gr: 0   COMPLEX 2   Preds:
Node: **Tuberculosis**
  gr: 1   COMPLEX 4   Preds: VisitToAsia
Node: **LungCancer**
  gr: 1   COMPLEX 4   Preds: Smoking
Node: **Bronchitis**
  gr: 1   COMPLEX 4   Preds: Smoking
Node: **TbOrCa**
  gr: 2   COMPLEX 8   Preds: Tuberculosis LungCancer
Node: **XRay**
  gr: 1   COMPLEX 4   Preds: TbOrCa
Node: **Dyspnea**
  gr: 2   COMPLEX 8   Preds: TbOrCa Bronchitis

```
................. 1 2 3 4 5 6 7 8
 1 VisitToAsia....0 0 1 0 0 0 0 0
 2 Smoking........0 0 0 1 1 0 0 0
 3 Tuberculosis...0 0 0 0 0 1 0 0
 4 LungCancer.....0 0 0 0 0 1 0 0
 5 Bronchitis.....0 0 0 0 0 0 0 1
 6 TbOrCa.........0 0 0 0 0 0 1 1
 7 XRay...........0 0 0 0 0 0 0 0
 8 Dyspnea........0 0 0 0 0 0 0 0
```

Maxszpot: 18
TOTAL COMPLEX: 36   MAX COMPLEX: 8

```
> summary.network( marnod.eval( asia), verbose=TRUE)
```

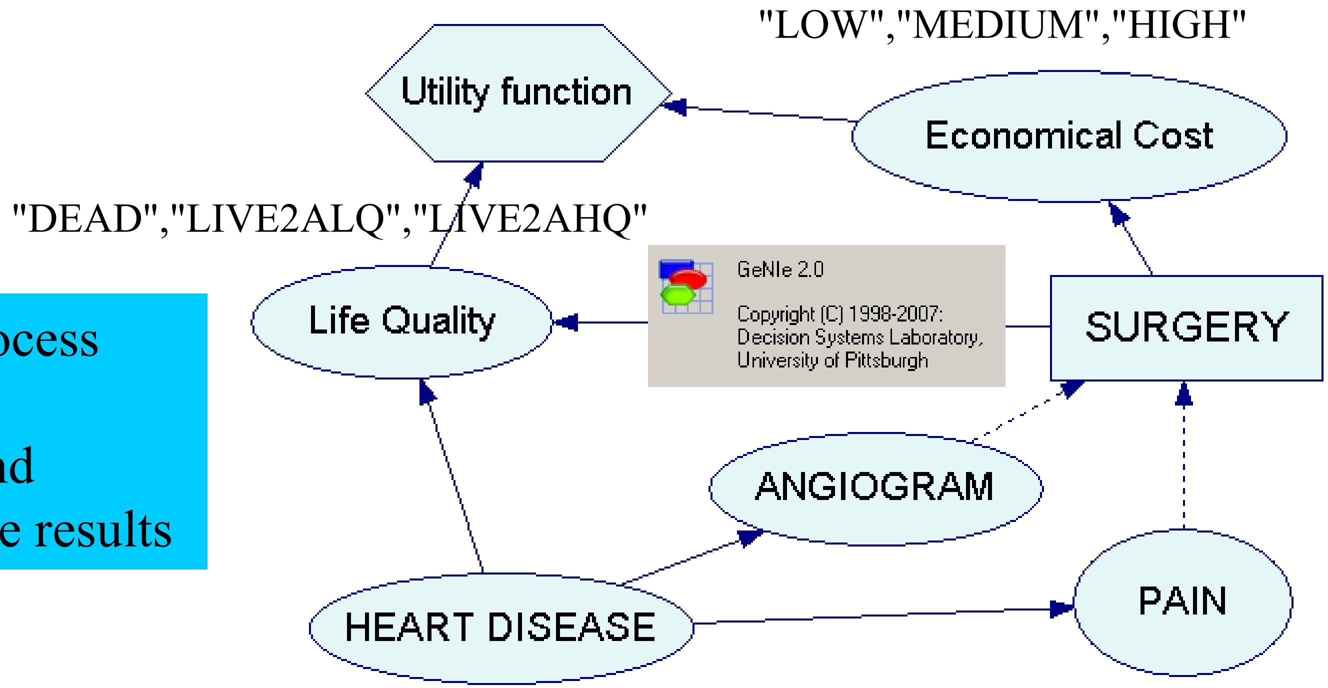
Network summary: netname   Nodes: 8
Arcs: 18   Max preds: 5   Max succs: 5

Node: **VisitToAsia**
  gr: 1   COMPLEX 4   Preds: Tuberculosis
Node: **Smoking**
  gr: 2   COMPLEX 8   Preds: LungCancer Bronchitis
Node: **Tuberculosis**
  gr: 4   COMPLEX 32   Preds: Bronchitis XRay Dyspnea TbOrCa
Node: **LungCancer**
  gr: 5   COMPLEX 64   Preds: Tuberculosis Bronchitis XRay Dyspnea TbOrCa
Node: **Bronchitis**
  gr: 3   COMPLEX 16   Preds: XRay Dyspnea TbOrCa
Node: **TbOrCa**
  gr: 0   COMPLEX 2   Preds:
Node: **XRay**
  gr: 2   COMPLEX 8   Preds: Dyspnea TbOrCa
Node: **Dyspnea**
  gr: 1   COMPLEX 4   Preds: TbOrCa

```
................. 1 2 3 4 5 6 7 8
 1 VisitToAsia....0 0 0 0 0 0 0 0
 2 Smoking........0 0 0 0 0 0 0 0
 3 Tuberculosis...1 0 0 1 0 0 0 0
 4 LungCancer.....0 1 0 0 0 0 0 0
 5 Bronchitis.....0 1 1 1 0 0 0 0
 6 TbOrCa.........0 0 1 1 1 0 1 1
 7 XRay...........0 0 1 1 1 0 0 0
 8 Dyspnea........0 0 1 1 1 0 1 0
```

Maxszpot: 114
TOTAL COMPLEX: 138   MAX COMPLEX: 64

Influence diagram **definition** and evaluation: ByPass



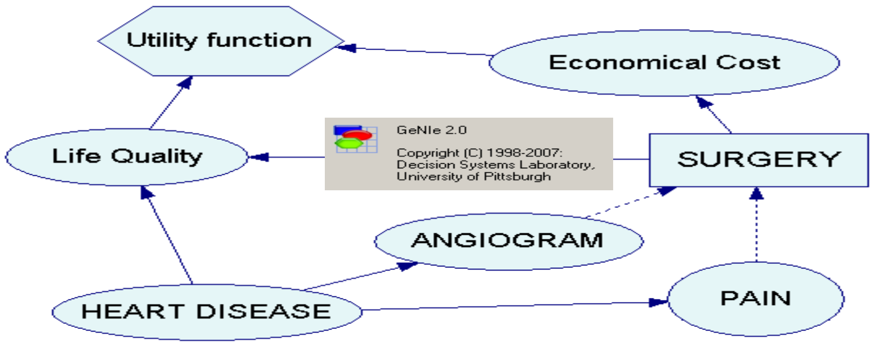
ID represents a decision process under uncertainty with a decision sequence and preferences (utility) over the results

```
•SURGERY = node( Type = "DECISION",
Name = "HEARTSURGERY", Values=c("NO","YES"),
Preds=c("PAIN","ANGIOGRAM"), Pots=matrix( data = c(1.0),
dimnames = list("phase","SURGERY"))),
```

```
•UTILITY = node(Type="UTILITY", Name="UTILITY", Values=c(0.0,1.0),
Preds=c("LIFEQ","ECONOMICALC"),
Pots=matrix( data=c(
1.0, 0.90, 0.70, 0.80, 0.50,
0.10, 1.40, 1.50, 1.80),
  nrow=9,ncol=1,byrow=TRUE, dimnames=list( NULL, c("UTILITY"))))
```

Influence diagram definition and **evaluation**: ByPass

Evaluation output is an optimal decision table for every decision




```
;Decision:  SURGERY
;Preds utility node:  UTILITY  < PAIN ANGIOGRAM SURGERY >
File:  dec-SURGERY ;
S:  10 SURGERY 2 ;
Val:  NO YES ;
Att:  200 PAIN 2 ;
Val:  ABSENT PRESENT ;
Att:  300 ANGIOGRAM 2 ;
Val:  NEGATIVE POSITIVE ;
Att:  400 SURGERY 2 ;
Val:  NO YES ;
```

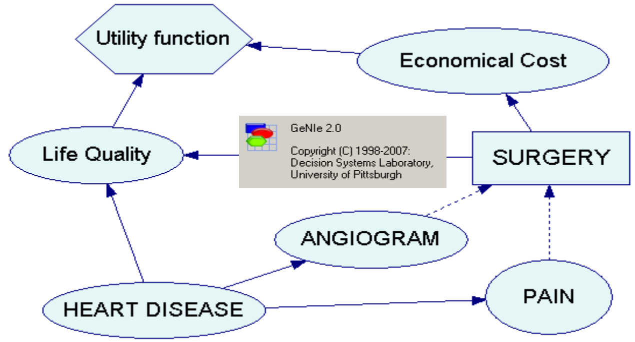
PAIN	**ANGIOGRAM**	**SURGERY**	**Utility**	
ABSENT	NEGATIVE	**NO**	**0.74673**	max
ABSENT	NEGATIVE	YES	0.64070	
ABSENT	POSITIVE	**NO**	**0.65233**	max
ABSENT	POSITIVE	YES	0.64598	
PRESENT	NEGATIVE	**NO**	**0.74453**	max
PRESENT	NEGATIVE	YES	0.64083	
PRESENT	POSITIVE	NO	**0.63965**	max
PRESENT	POSITIVE	**YES**	0.64668	

Optimal policy:

Pain Absent & Angiogram Negative then Surgery No
Pain Absent & Angiogram Positive then Surgery No
Pain Present & Angiogram Negative then Surgery No
Pain Present & Angiogram Positive then Surgery Yes

Influence diagram definition and evaluation: ByPass; **explanation**

KBM2L: Kowledge Base Matrix to List




```
;Decision:  SURGERY
;Preds utility node:  UTILITY  < PAIN ANGIOGRAM SURGERY >
File:  dec-SURGERY ;
S:  10 SURGERY 2 ;
Val:  NO YES ;
Att:  200 PAIN 2 ;
Val:  ABSENT PRESENT ;
Att:  300 ANGIOGRAM 2 ;
Val:  NEGATIVE POSITIVE ;
Att:  400 SURGERY 2 ;
Val:  NO YES ;
```

PAIN	ANGIOGRAM	SURGERY	Utility
ABSENT	NEGATIVE	**NO**	0.746733
ABSENT	NEGATIVE	YES	0.640708
ABSENT	POSITIVE	**NO**	0.652331
ABSENT	POSITIVE	YES	0.645981
PRESENT	NEGATIVE	**NO**	0.744533
PRESENT	NEGATIVE	YES	0.640831
PRESENT	POSITIVE	NO	0.639655
PRESENT	POSITIVE	**YES**	0.646689

Table – Multidimensional Matrix

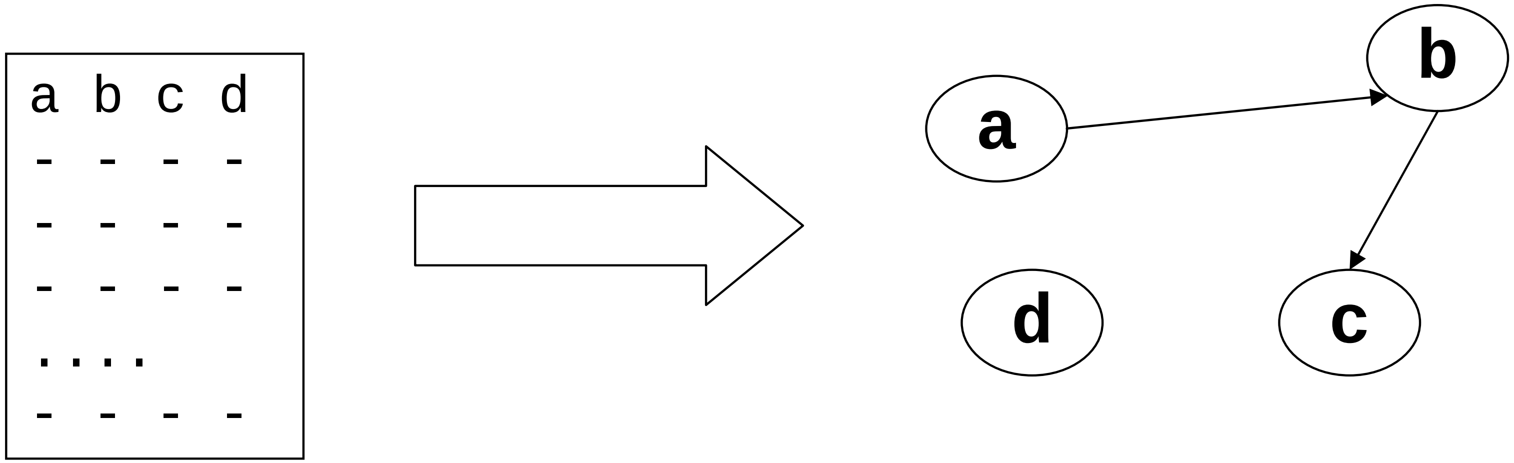
KBM2L: <(Present,Negative),No| <(Present,Positive),Yes|

Explanation: Surgery No <- (Pain Absent) OR (Angiogram Negative)
Sugery Yes <- (Pain Present) AND (Angiogram Positive)

The best explanation is available using the most concise list; How?
Searching the proper permutation of the attibutes (and domains) on the table!
Also useful for conditional probability tables.

LBN

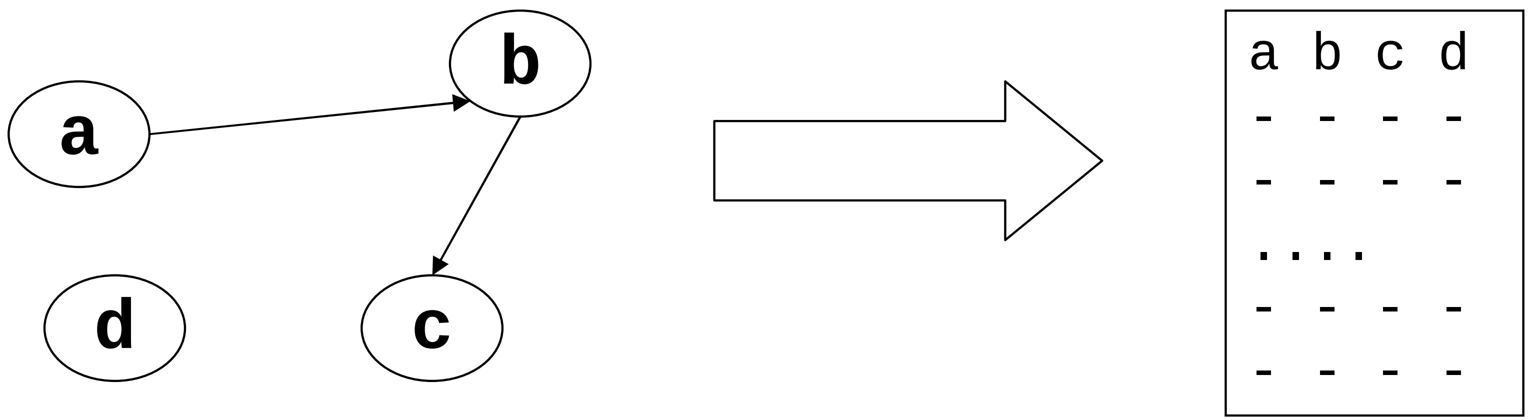
Learning a marginal / naive / generic Bayesian network model for Classification



Estructure learning and probability model estimation

SBN

Simulating a data set from the model for Inference



Sampling the model P(abcd)=P(a)P(d)P(b|a)P(c|b)

Future lines of research:

•more general decision networks (continuous variables, several utility nodes, non sequential decision nodes,. . .)

•alternatives to the (large) conditional probability tables (linear models) and utility tables (multiattribute utility functions)

•implementation of an R package for KBM2L

•evaluation and learning algorithms from data

•complex queries, MPE, MAP

We are interesting on paralell evaluation of huge models, using packages like *snow*, i.e. very large decision sequences

Departamento Inteligencia Artificial

POLITÉCNICA

References:

•Bielza, C., Fernández del Pozo, J.A., Lucas, P. (2008).
Explaining clinical decisions by extracting regularity patterns,
Decision Support Systems, 44, 397--408.

•Fernández del Pozo, J.A. and Bielza, C. and Gómez, M. (2005).
A List-Based Compact Representation for Large Decision Tables Management.
European Journal of Operational Research, 160, 638--662.

•E. Castillo, J. Gutierrez, and A. Hadi. (1997).
Expert Systems and Probabilistic Network Models.
Springer-Verlag, first edition.

•Naive links to References :
http://en.wikipedia.org/wiki/Bayesian_network
http://en.wikipedia.org/wiki/Influence_diagram
http://en.wikipedia.org/wiki/Graphical_model