CONSTRUCTING EXPLANATORY PROGNOSTIC PROFILES FROM CONDITIONAL PROBABILITY TABLES BY CLUSTER-KBM2L ANALYSIS

Universidad Politécnica de Madrid, Spain
Radboud University Nijmegen, The Netherlands

J.A. Fernandez del Pozo[a], C. Bielza[a] and Peter J.F. Lucas[b]

13th Biennial European Meeting of the SMDM, May 30-June 2, 2010

PROGNOSIS PROFILES//CLUSTER-KBM2L ANALYSIS

BACKGROUND: Clinical prognosis, Conditional probability

METHODS: Graphical models, Clustering CPTs, KBM2L synthesis

RESULTS: Prognosis variables in NHL model

CONCLUSION: Qualitative description of probabilistic models

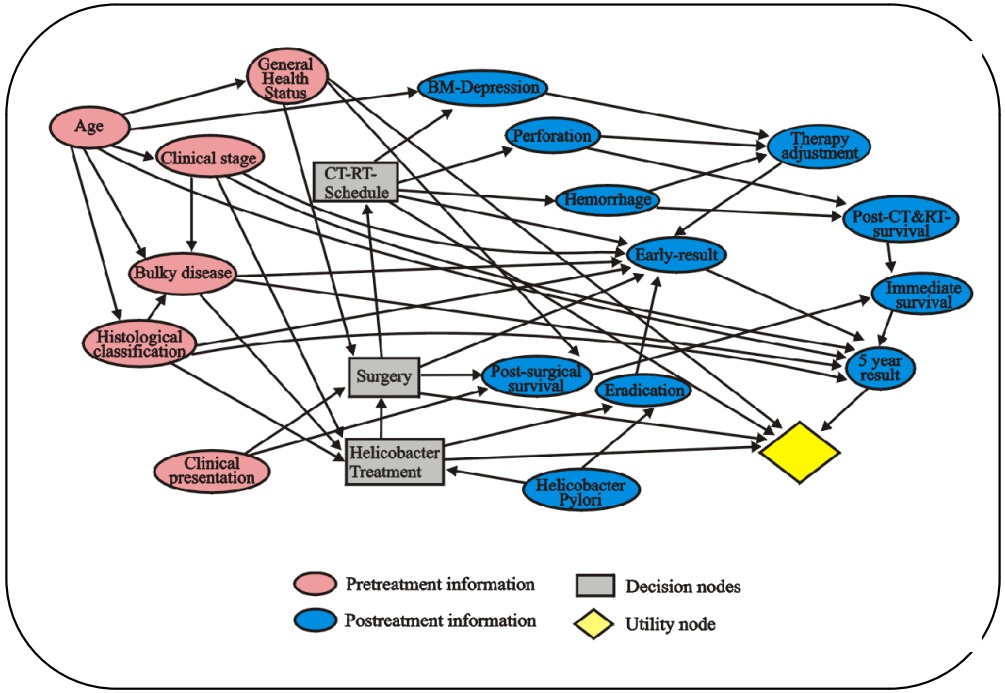
Clinical decision making: diagnosis and prognosis

A prediction model for clinical prognosis may act as a source for clinical decision making; when based on probability theory it will typically include random variables dependent of many factors.

/———/

5% of gastric tumors. Chronic infection Helicobacter pylori. The model can make a diagnosis and suggest a treatment.

Decision model of Primary gastric non-Hodgkin lymphoma[a], *gastric NHL influence diagram*



[a]Lucas et al., 1998, Methods of Information in Medicine

Conditional probability

17 chance nodes (ellipses), 1 value node (diamond) and 3 decision nodes (rectangles). 42 arcs, **8,282 probability entries** and 144 utility entries

Table: Gastric NHL variables with their possible values

Variable	Possible values	Variable	Possible values
helicobacter-treatment	No, Yes	age	v10.19, v20.29, ...
surgery	None, Curative,		..., v80.89, GE90
	Palliative	eradication	No, Yes
ct-rt-schedule	None, Radio,	bone.marrow.depression	No, Yes
	Chemo, Ch.Next.Rad	perforation	No, Yes
		hemorrhage	No, Yes
general-health-status	Poor, Average, Good	therapy.adjustment	No, Yes
clinical-stage	I, II1, II2, III, IV	post.ct-rt.survival	No, Yes
bulky-disease	Yes, No	post.surgical.survival	No, Yes
histological-classification	Low.Grade,	immediate.survival	No, Yes
	High.Grade	EARLY.RESULT	CR (complete remission),
helicobacter-pylori	Absent, Present		PR (partial remission),
clinical-presentation	None, Hemorrhage,		NC (no change),
	Perforation,		PD (progressive disease)
	Obstruction	FIVE.YEAR.RESULT	alive, dead

Conditional probability tables

A conditional probability table, or CPT for short, is a table with probabilities for a discrete random variable X conditioned on a set of other discrete random variables Y.

One position (i, j) in the CPT yields $P(X = x_j | Y = y_i)$.

Example of a probabilistic relationship among clinical variables:

$X \equiv$ POST-ChemoTherapy-RadioTherapy-SURVIVAL (No/Yes),
$Y_1 \equiv$ Perforation (No/Yes) and $Y_2 \equiv$ Hemorrarhage (No/Yes).

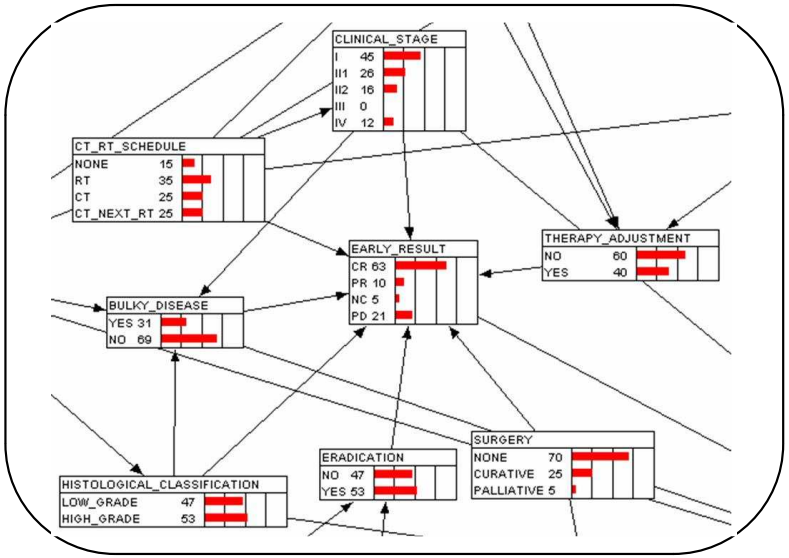
Table: POST-ChemoTherapy-RadioTherapy-SURVIVAL CPT

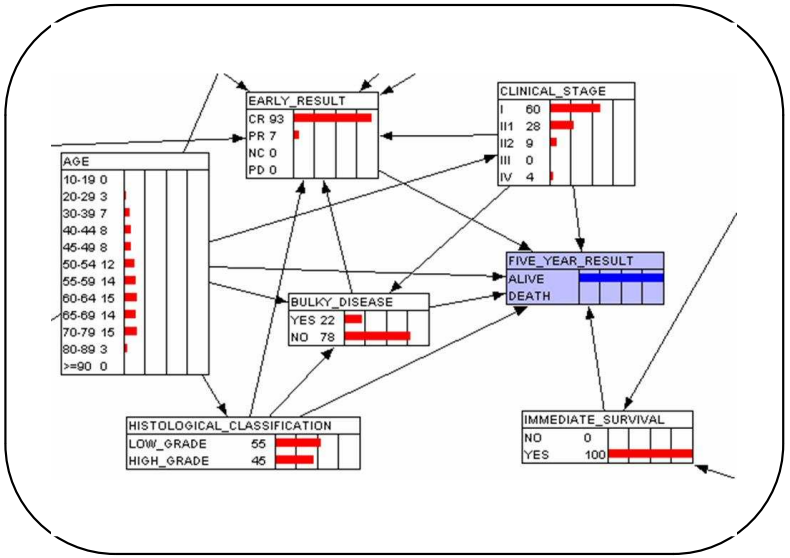
X	Y_1	Y_2	$P(X\|Y_1, Y_2)$	X	Y_1	Y_2	$P(X\|Y_1, Y_2)$
No	No	No	0.00	No	Yes	No	0.20
Yes	No	No	1.00	Yes	Yes	No	0.80
No	No	Yes	0.10	No	Yes	Yes	0.25
Yes	No	Yes	0.90	Yes	Yes	Yes	0.75

Hypotesis: there are two classes of Post-Survivor patients according to $\{(0.0, 1.0), (0.1, 0.9)\}$ and $\{(0.2, 0.80), (0.25, 0.75)\}$

Prognostic probabilistic models

CPTs of prognostic probabilistic models may be very large and complex, and, as a consequence, it may be hard to grasp the tables' content when building or using a model.





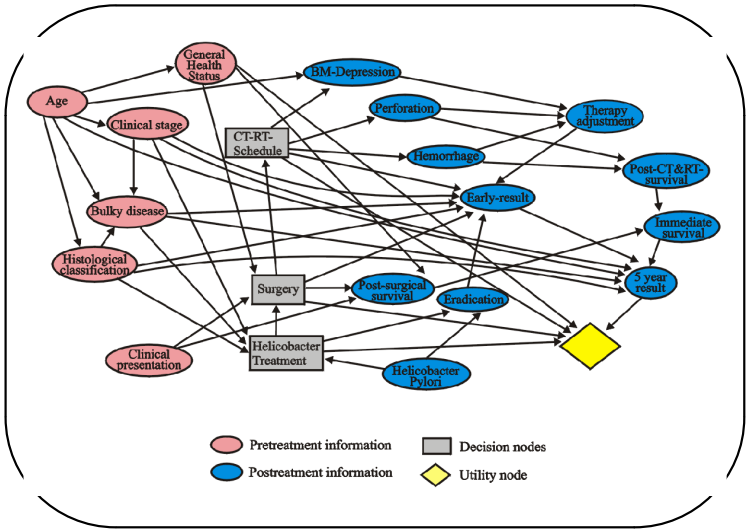
Yet, a good understanding of their content is required when developing a model that is accurate.

Probabilistic graphical models

Our clinical knowledge framework under uncertainty is a modern and useful decision-theoretic model: an **influence diagram** (Shachter, 1986).

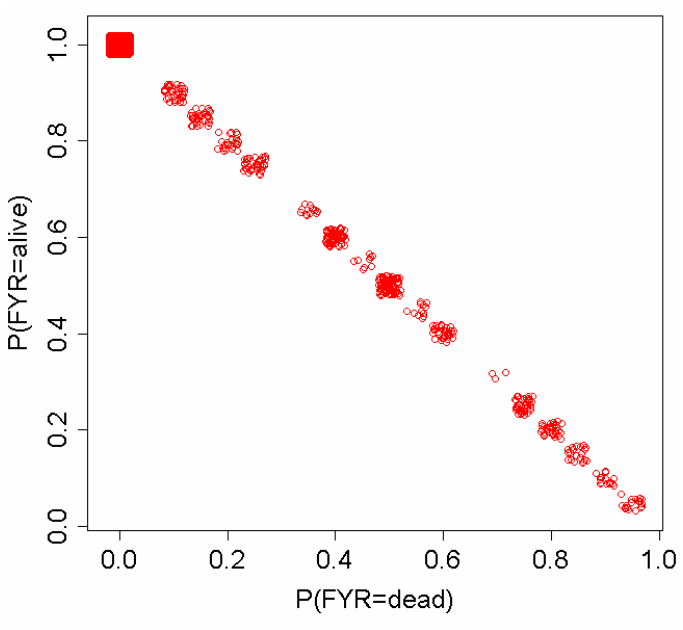
It consists of an acyclic directed graph with associated **probabilities (CPTs)** and utilities, respectively modeling the uncertainties and preferences tied in with the stated problem.

Influence diagram evaluation outputs a set of **optimal decision tables** as the optimal decision policies support and a set of **a posteriori conditional probability tables** as diagnosis inference results.



Clustering to label the FIVE.YEAR.RESULT (FYR) CPT

Medoids ($P(FYR = dead)$, $P(FYR = alive)$) using four clusters: $(0.0, 1.0)$, $(0.8, 0.2)$, $(0.5, 0.5)$, $(0.15, 0.85)$



The profiles may be labeled as: *Grim, Favourable, Unfavourable and Bad*.

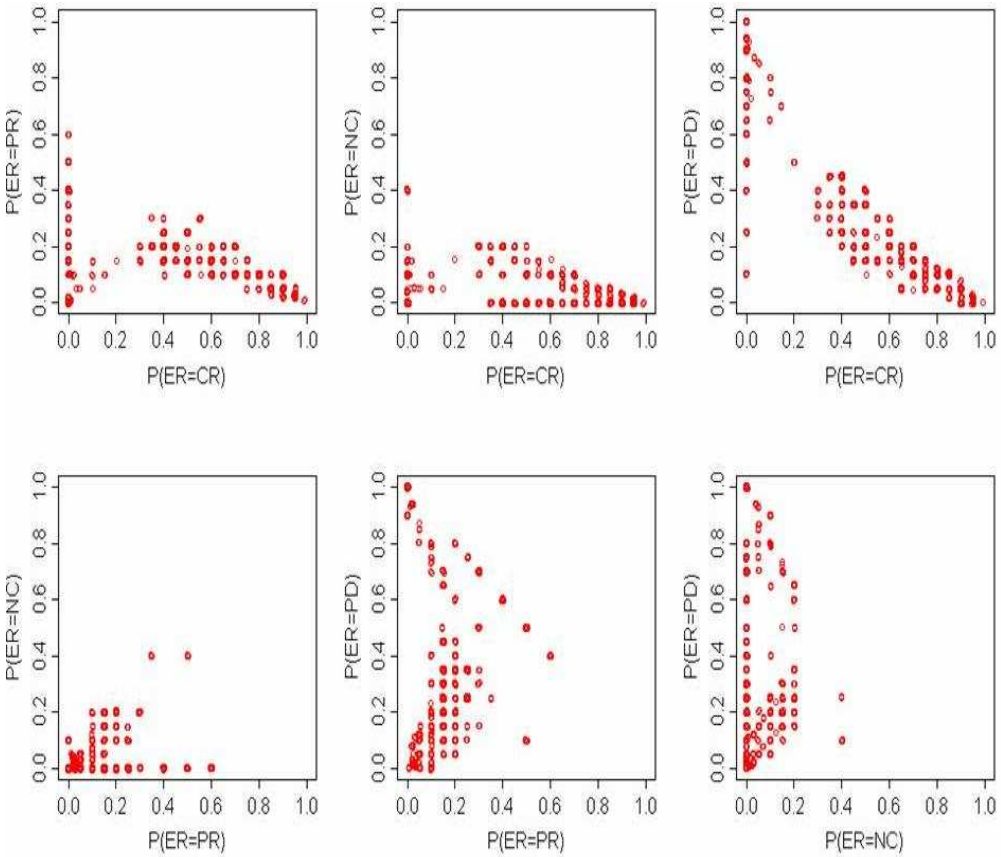
Similarity is measured via the **PAM** (Partitioning Around Medoids) cluster algorithm merging the CPT rows into k clusters).

The resulting cluster *representatives*, here medoids, with associated cluster profiles were optimized using the **KBM2L** procedure, resulting in explanations of the profiles.

Expert doctors propose how many clusters fit the prognosis variables.

Clustering to label the EARLY.RESULT (ER) CPT

Medoids ($P(ER = CR)$, $P(ER = PR)$, $P(ER = NC)$, $P(ER = PD)$) using four clusters: $(0.0, 0.0, 0.0, 1.0)$, $(0.8, 0.1, 0.0, 0.1)$, $(0.45, 0.2, 0.0, 0.35)$, $(0.0, 0.3, 0.0, 0.7)$



We take into acount the diference among medoids and the similarity into the cluster.

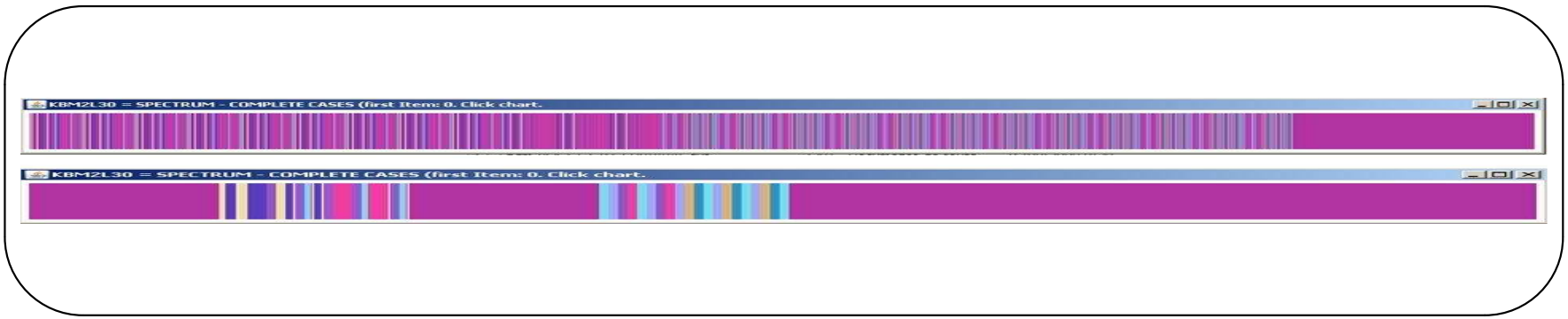
With four dimensions (or more) on data we need the expert support to fix the number of clusters k.

KBM2L synthesis

We applied a procedure for the CPT analysis using the recently proposed KBM2L framework, Fernańdez del Pozo et al (2005), to the gastric NHL influence diagram. **KBM2L provides a compact list composed of items** $< index; response|$**.**

Response $\rightarrow X$ conditional probability distribution, labeled by the cluster. **Index** $\rightarrow Y$ values from the last row in any group of consecutive rows in the CPT with similar response. KBM2L optimal order of $\{Y_i\}$ that yields more similarity, more coalesced CPTs.

Spectrum chart of the probability distribution clusters.



We analyze two CPTs

We analyzed two CPTs associated with variables Early Result (ER) and FiveYears Result (FYR) consisting of 970 and 1920 conditional probability distributions, respectively.

For complex **variable FYR (alive, death)**, Y consisted of six variables: Age, Clinical Stage, Early Result, Inmediate Survival Histological Classification and Bulky Disease.

For complex **variable ER (CR, PR, NC, PD)**, Y consisted of seven variables: Therapy adjustement, Erradication, Clinical Stage, Bulky Disease, Histological Classification, ChemoTherapy-RadioTherapy Schedule and Surgery.

We also synthesized its CPT into up to four clusters according to the physician's preference. The following four clusters were obtained:

Variable FYR (alive, death)

Prognosis profile/ Probability	Description/ Explanation
Favourable $P(FYR = alive) > 0.8$	full, unadjusted treatment and gastroscopically confirmed complete or partial remission of NHL.
Unfavourable $0.35 < P(FYR = alive) < 0.60$	full, unadjusted treatment, gastroscopically confirmed complete or partial remission, bulky disease with low-grade histology or stage I NHL.
Bad $P(FYR = alive) < 0.25$	gastroscopically confirmed partial remission after unadjusted treatment.
Grim $P(FYR = alive) = 0.0$	the side effects of treatment with associated reduced effectiveness of adjusted treatment, and gastroscopically confirmed disease progression or no change.

Variable ER (CR, PR, NC, PD)

Prognosis profile/ Probability	Description/ Explanation
Favourable $P(ER = CR) > 0.6$ $P(ER = CR\ or\ PR) > 0.7$	.
Unfavourable $0.35 < P(ER = CR\ or\ PR) < 0.8$, $0.2 < P(ER = NC\ or\ PD) < 0.65$	.
Bad $0.2 < P(ER = PR) < 0.6$, $0.4 < P(ER = PD) < 0.8$	.
Grim $P(ER = PD) > 0.8$	.

To construct, validate and explain large CPTs

- Profiles of conditional probability distributions $\sim$ clusters and medoids.
- Clinical state of art: the literature states which profiles are used. The number of clusters are pointed out by the clinical guidelines, but multimodality among distributions suggests more clusters to arise more homogeneous cluster.
- Better explanations: less items with short explanations. Explanation trees for complex profiles. Explanation and Relevance of variables.
- This technique was shown to be promising for assisting clinical researchers in constructing, validating and explaining clinical models with large CPTs.

References I

Shachter, R.D.:
Evaluating Influence Diagrams.
Operations Research, 34 6 (1986) 871-882

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

Lucas, P., Boot, H., Taal, B.:
Computer-Based Decision-Support in the Management of Primary Gastric non-Hodgkin Lymphoma.
Methods of Information in Medicine, 37 (1998) 206-219