Rovira i Virgili University

Research Group on Artificial Intelligence
(Banzai)

Repository of background knowledge

Authors:
Joan Albert López Vallverdú
Dr. David Riaño

September 7, 2011

Contents

Chapter 1

Introduction

The datasets used to build this repository have been provided by the UCI Repository of Machine Learning[1]. Table 1.1 contains a brief summary about the main characteristics of the datasets of each domain.

Table 1.1: Summary table of the datasets of the domains tested

	Instances	Questions	Decisions
Diabetes	768	8	2
Heart disease	303	13	2
Post-operative	90	8	3
Thyroid	3772	20	3

The medical background knowledge for each of the domains has been generated with the collaboration of the Hospital Clinic (Barcelona, Catalonia, Spain) and the Hospital Consortium SAGESSA (Reus, Catalonia, Spain).

The notation used in the following sections is:

- K_c : Cost functions for criteria on the questions. They are defined as $K_c : Q \rightarrow [0..1]$ where Q is the set of questions and c is a letter representing a criterion (e - economic cost, h - health risk, m - script or adherence to medical standards, t - decision time and c - comfortability)

- $\leq_c$: Partial orders for criteria on the questions, defined over the set of questions Q.

[1] http://archive.ics.uci.edu/ml/

- Ie_c : Cost functions for criteria on the decisions (type I errors). They are defined as $Ie_c : D \rightarrow [0..1]$ where D is the set of decisions.
- $\leq_{Ie-c}$: Partial orders for criteria on the decisions (type I errors), defined over the set of decisions D.
- IIe_c : Cost functions for criteria on the decisions (type II errors). They are defined as $IIe_c : D \rightarrow [0..1]$ where D is the set of decisions.
- $\leq_{IIe-c}$: Partial orders for criteria on the decisions (type II errors), defined over the set of decisions D.

Chapter 2

Diabetes domain

In this domain we have to decide whether a patient is tested positive for diabetes or not. The questions used are described in table 2.1.

Table 2.1: Description of the questions of the diabetes domain

	Description	*Values*
$times - pregnant$	Number of times pregnant	integer
$plasma - glucose$	Plasma glucose in an oral glucose tolerance test	integer
$diastolic - blood$	Diastolic blood pressure (mm Hg)	integer
$triceps - skin$	Triceps skin fold thickness (mm)	integer
$serum - insulin$	2-Hour serum insulin (mu U/ml)	integer
$body - mass$	Body mass index (weight in kg/(height in m)^2)	real
$diabetes - pedigree$	Diabetes pedigree function	real
age	Age (years)	integer

The possible final decisions are described in table 2.2.

Table 2.2: Description of the decisions of the diabetes domain

	Description
0	Tested negative for diabetes
1	Tested positive for diabetes

In this domain each question is provided by a different test (see table 2.3).

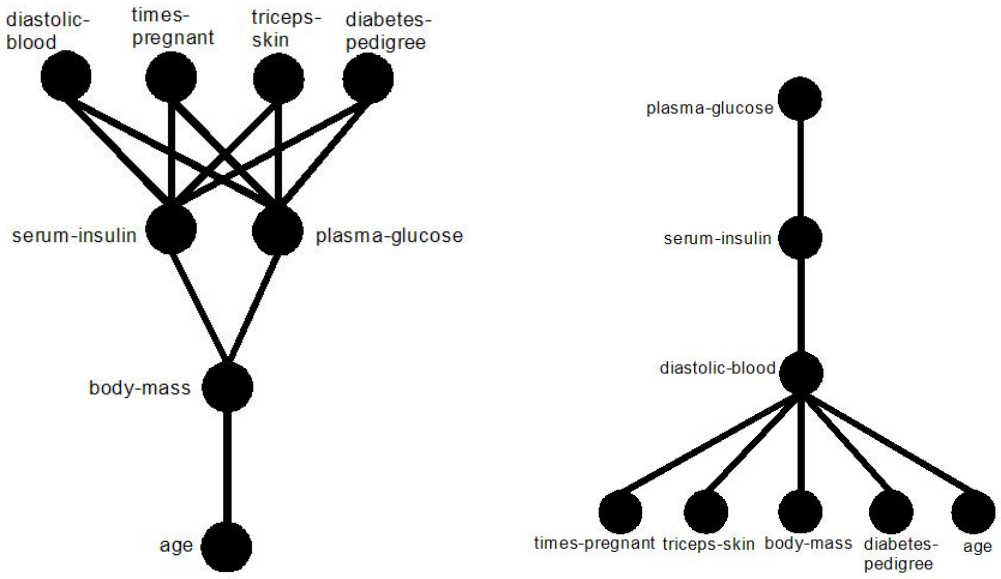
For criteria on the questions, the final values for the cost functions K_e and K_t are detailed in table 2.4 and the final partial orders $\leq_m$ and $\leq_c$ are depicted in figure 2.1. We do not show the partial order $\leq_h$ because of the lack of risked tests in this domain.

Table 2.3: Tests needed for each question in the diabetes domain

	α	β	γ	δ	ϵ	ζ	η	θ
$times - pregnant$	$\times$							
$plasma - glucose$		$\times$						
$diastolic - blood$			$\times$					
$triceps - skin$				$\times$				
$serum - insulin$					$\times$			
$body - mass$						$\times$		
$diabetes - pedigree$							$\times$	
age								$\times$

Table 2.4: The values of K_x for the diabetes domain

	K_e	K_t
$times - pregnant$	0	0
$plasma - glucose$	0.5	0.5
$diastolic - blood$	0.25	0.008
$triceps - skin$	0	0.008
$serum - insulin$	1	1
$body - mass$	0	0.017
$diabetes - pedigree$	0	0
age	0	0



(a) $\leq_m$ (b) $\leq_c$

Figure 2.1: The partial orders $\leq_x$ for the diabetes domain

For type I error criteria we use cost functions which are detailed in table 2.5.

Table 2.5: The values of Ie_x for the diabetes domain

	0	1
Ie_e	0.4	1
Ie_t	0	0
Ie_h	1	0.4
Ie_c	0	1

For type II error criteria, we also represent the knowledge using cost functions depicted in table 2.6.

Table 2.6: The values of IIe_h for the diabetes domain

	0	1
IIe_h	0.8	1

Chapter 3

Heart disease domain

The aim of this domain is to detect the presence of heart disease in the patient. The questions used are described in table 3.1.

Table 3.1: Description of the questions of the heart disease domain

	Description	*Values*
age	Age in years	integer
sex	Sex	1=male, 0=female
cp	Chest pain type	1=typical angina, 2=atypical angina, 3=non-anginal, 4=asymptomatic
trestbps	Resting blood pressure	integer
chol	Serum cholesterol in mg/dl	integer
fbs	Fasting blood sugar > 120 mg/dl?	1=true, 0=false
restecg	Resting electrocard. results	0=normal, 1=ST-T wave, 2=left vent. hypertrophy
thalach	Maximum heart rate achieved	integer
exang	Exercise induced angina	1=yes, 0=no
oldpeak	ST depression induced by exercise relative to rest	real
slope	Slope of the peak exercise ST segment	1=upsloping, 2=flat, 3=downsloping
ca	Number of major vessels	integer between 0 and 3
thal	Thal	3=normal, 6=fixed defect, 7=reversable defect

The possible final decisions are described in table 3.2.

The grouping of the questions in tests is shown in table 3.3.

For criteria on the questions, the final values for the cost functions K_e and K_t are detailed in table 3.4 and the final partial orders $\leq_m$, $\leq_h$ and

Table 3.2: Description of the decisions of the heart disease domain

	Description
0	$< 50\%$ diameter narrowing (angiographic disease status)
1	$> 50\%$ diameter narrowing (angiographic disease status)

Table 3.3: Tests needed for each question in the diabetes domain

	α	β	γ	δ	ϵ	ζ	η	θ
age	$\times$							
sex		$\times$						
cp			$\times$					
trestbps				$\times$				
chol					$\times$			
fbs					$\times$			
restecg						$\times$		
thalach							$\times$	
exang							$\times$	
oldpeak							$\times$	
slope							$\times$	
ca								$\times$
thal								$\times$

$\leq_c$ are depicted in figure 3.1. In this case the partial orders $\leq_h$ and $\leq_c$ are identical.

Table 3.4: The values of K_x for the heart disease domain

	K_e	K_t
age	0	0
sex	0	0
cp	0	0
$trestbps$	0	0.008
$chol$	0.063	1
fbs	0.063	1
$restecg$	0.041	0.042
$thalach$	0.083	0.25
$exang$	0.083	0.25
$oldpeak$	0.083	0.25
$slope$	0.083	0.25
ca	1	0.5
$thal$	1	0.5

For type I error criteria, the values for the cost function Ie_t are detailed in table 3.5.

Table 3.5: The values of Ie_x for the heart disease domain

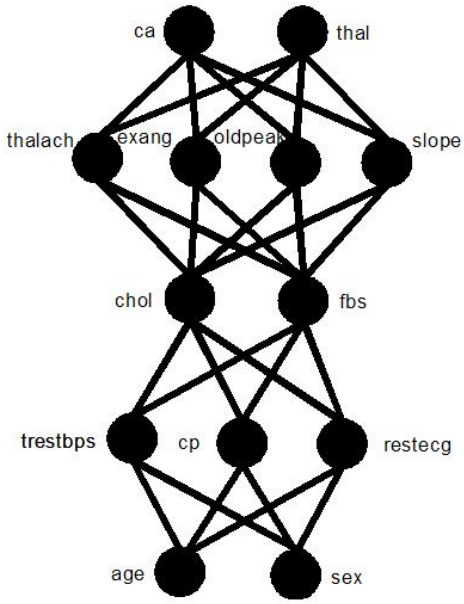
	0	1
Ie_t	0	0

The partial orders $\leq_{Ie-e}$, $\leq_{Ie-h}$ and $\leq_{Ie-c}$ for type I error criteria are identical. They are depicted in figure 3.2.

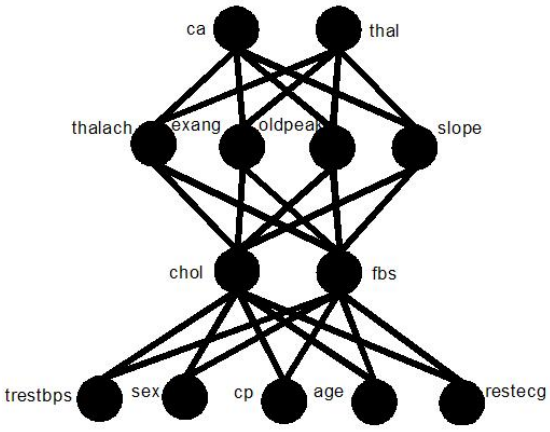
For type II error criteria, we represent the knowledge using a cost function IIe_h depicted table 3.6.

Table 3.6: The values of IIe_h for the heart disease domain

	0	1
IIe_h	0.25	1



(a) $\leq_m$



(b) $\leq_h$ / $\leq_c$

Figure 3.1: The partial orders $\leq_x$ for the heart disease domain

Figure 3.2: The partial order $\leq_{Ie-e}$ / $\leq_{Ie-h}$ / $\leq_{Ie-c}$ for the heart disease domain

Chapter 4

Post-operative domain

In this domain we deal with the decision problem of determining where the patients in a post-operative recovery area should be sent to next. The questions used are described in table 4.1.

Table 4.1: Description of the questions of the post-operative domain

	Description	*Values*
$L-CORE$	Patient's internal temperature	high, mid, low
$L-SURF$	Patient's surface temperature	high, mid, low
$L-O2$	Oxygen saturation	excellent, good, fair, poor
$L-BP$	Last measurement of blood pressure	high, mid, low
$SURF-STBL$	Stability of patient's surface temperature	stable, mod-stable, unstable
$CORE-STBL$	Stability of patient's core temperature	stable, mod-stable, unstable
$BP-STBL$	Stability of patient's blood pressure	stable, mod-stable, unstable
$COMFORT$	Patient's perceived comfort at discharge	integer between 0 and 20

The possible final decisions are described in table 4.2.

Table 4.2: Description of the decisions of the post-operative domain

	Description
I	Patient sent to Intensive Care Unit
S	Patient prepared to go home
A	Patient sent to general hospital floor

According to the background knowledge obtained from the physicians, the questions are grouped in the six tests of table 4.3.

Table 4.3: Tests needed for each question in the post-operative domain

	α	β	γ	δ	ϵ	ζ
$L-CORE$	$\times$					
$L-SURF$	$\times$					
$L-O2$		$\times$				
$L-BP$			$\times$			
$SURF-STBL$				$\times$		
$CORE-STBL$				$\times$		
$BP-STBL$					$\times$	
$COMFORT$						$\times$

For criteria on the questions, the final values for the cost functions K_e and K_t are detailed in table 4.4 and the final partial orders $\leq_m$ and $\leq_c$ are depicted in figure 4.1. As in the diabetes domain, the partial order $\leq_h$ is not shown because in this domain there not seem to be tests more risked than others.

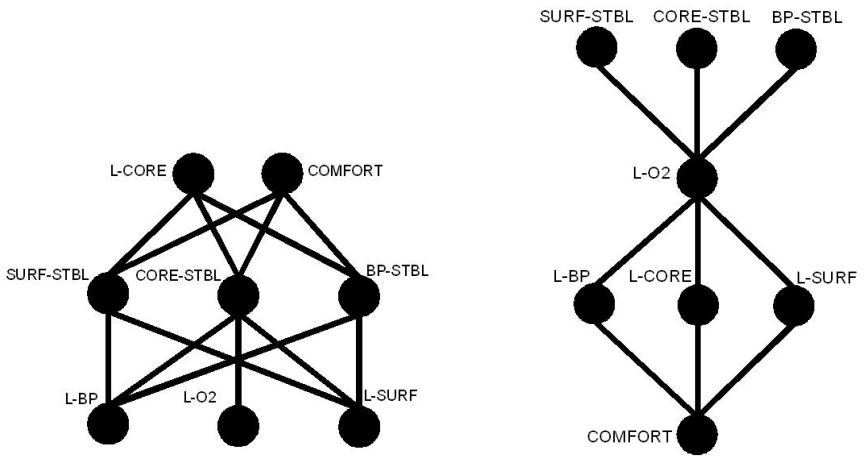
Table 4.4: The values of K_x for the post-operative domain

	K_e	K_t
$L-CORE$	0	0.02
$L-SURF$	0	0.02
$L-O2$	1	0
$L-BP$	1	0.03
$SURF-STBL$	0	1
$CORE-STBL$	0	1
$BP-STBL$	0	1
$COMFORT$	0	0

For type I error criteria, the values for the cost functions Ie_e and Ie_t are detailed in table 4.5.

The partial orders $\leq_{Ie-h}$ and $\leq_{Ie-c}$ corresponding to type I error criteria are depicted in figure 4.2.

For type II error criteria, the final partial order $\leq_{IIe-h}$ is depicted in figure 4.3.

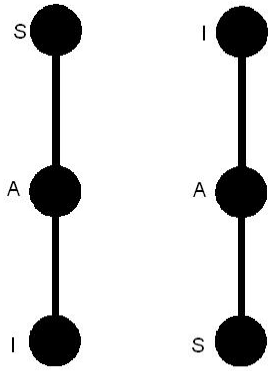


(a) $\leq_m$ (b) $\leq_c$

Figure 4.1: The partial orders $\leq_x$ for the post-operative domain

Table 4.5: The values of Ie_x for the post-operative domain

	I	S	A
Ie_e	0.14	1	0.29
Ie_t	1	0	0.33



(a) $\leq_{Ie-h}$ (b) $\leq_{Ie-c}$

Figure 4.2: The partial orders $\leq_{Ie-x}$ for the post-operative domain

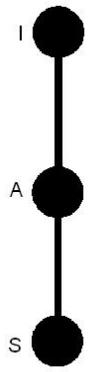


Figure 4.3: The partial order $\leq_{IIe-h}$ for the post-operative domain

Chapter 5

Thyroid domain

In this domain the patients are classified according to their thyroid functioning into three classes: normal (not hypothyroid), hyperfunction or subnormal functioning. The questions used are described in table 5.1.

Table 5.1: Description of the questions of the thyroid domain

	Description	*Values*
age	Age normalized	real
sex	Sex	1=male, 0=female
on_thyroxine	On thyroxine	0=false, 1=true
query_on_thyroxine	Query on thyroxine	0=false, 1=true
on_antithyroid_medication	On antithyroid medication	0=false, 1=true
sick	Sick	0=false, 1=true
pregnant	Pregnant	0=false, 1=true
thyroid_surgery	Thyroid surgery	0=false, 1=true
*I*131*_treatment*	I131 treatment	0=false, 1=true
query_hypothyroid	Query hypothyroid	0=false, 1=true
query_hyperthyroid	Query hyperthyroid	0=false, 1=true
lithium	Lithium	0=false, 1=true
goitre	Goitre	0=false, 1=true
tumor	Tumor	0=false, 1=true
hypopituitary	Hypopituitary	0=false, 1=true
psych	Psychological symptoms	0=false, 1=true
TSH	Thyroid-stimulating hormone	real
*T*3	Triiodothyronin	real
*TT*4	Total thyroxine	real
*T*4*U*	T4 uptake	real
FTI	Free thyroxine index	real

The possible final decisions are described in table 5.2.

Table 5.2: Description of the decisions of the heart disease domain

	Description
1	normal (not hypothyroid)
2	hyperfunction
3	subnormal functioning

Only a group of 5 questions are obtained with the same test (see table 5.3).

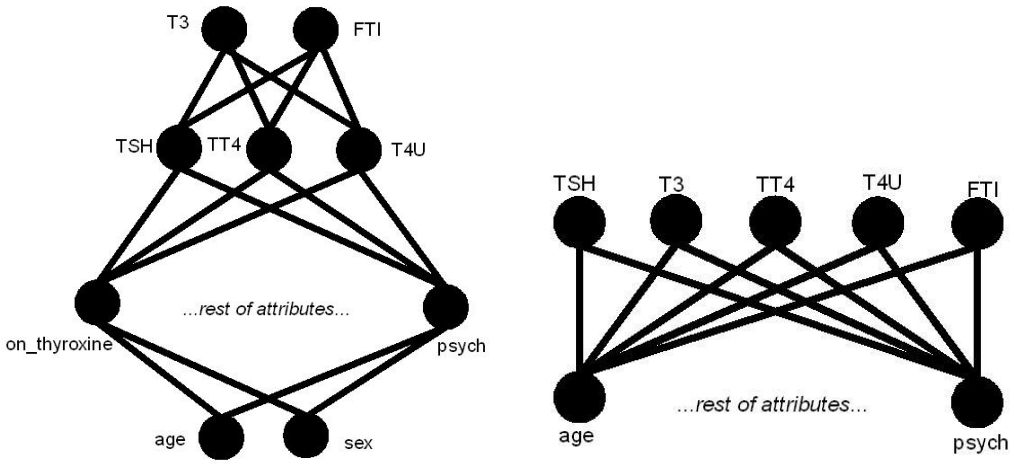
Table 5.3: Tests needed for each question in the diabetes domain

	α	β	γ	δ	ϵ	ζ	η	θ	ι	κ	λ	μ	ν	ξ	o	π	ρ
age	×																
sex		×															
on_thyroxine			×														
query_on_thyroxine				×													
on_antithyroid_medication					×												
sick						×											
pregnant							×										
thyroid_surgery								×									
*I*131*_treatment*									×								
query_hypothyroid										×							
query_hyperthyroid											×						
lithium												×					
goitre													×				
tumor														×			
hypopituitary															×		
psych																×	
TSH																	×
*T*3																	×
*TT*4																	×
*T*4*U*																	×
FTI																	×

For criteria on the questions we have only used partial orders. The partial orders $\leq_e$, $\leq_t$ and $\leq_c$ are the same and they are depicted in figure 5.1 beside the partial order $\leq_m$. For space reasons we have not included all the questions in the figure. We do not show the partial order $\leq_h$ because there are no risked tests.

For type I error criteria we only use cost functions (see table 5.4).

For type II error criteria, we represent the knowledge using a cost function IIe_h depicted table 5.5.



(a) $\leq_m$ (b) $\leq_e$ / $\leq_t$ / $\leq_c$

Figure 5.1: The partial orders $\leq_x$ for the thyroid domain

Table 5.4: The values of Ie_x for the thyroid domain

	1	2	3
Ie_e	0	1	1
Ie_t	1	1	1
Ie_h	1	1	1
Ie_c	0.5	1	1

Table 5.5: The values of IIe_h for the thyroid domain

	1	2	3
IIe_h	1	1	1