Application with Strata Decision Trees in SODAS 1.04

Instituto Nacional de Estadistica-Lisboa
26th January 2004

María del Carmen Bravo
Universidad Complutense de Madrid
carmen@sim.ucm.es

Summary

- Input / output
- Application data
- Application output:
  - Decisional node
  - Strata Description
- Prediction
- Visualisation and results
- Other applications
- Conclusion

Input and Output

- Input: Symbolic data
  - More complex and rich data:

    In SDT: Modal probabilistic data (probability distributions over a set of categories)
- Method: Segmentation trees for stratified data
- Output: Symbolic objects (concepts)
  - Represent an intention
  - Generally represented by a conjunction of properties
  - Their extension is the set of individuals which *are related* with the intention

Input and Output

- Some symbolic data advantages:
  - Aggregated data representation
  - Confidentiality preservation
  - Data volume reduction
- Symbolic object

  = Intention

  Symbolic description + recognition function of the extension

  + Extension

  { Individuals represented by the concept }

Example : [Terr = brittany] $\wedge$ [Sex ~ (Man(0.8), Woman(0.2))] $\wedge$ [RACT1 $\in$ {Agriculture, cattle,fishing}] $\wedge$ [Salary $\in$[1.2,3.1]]

Input Data. Symbolic data

- Input data units are divided into classes (2 in SDT) --> Z class variable
- Predictors (nominal/ modal probabilistic variables) --> $Y_1, \ldots, Y_n$
- Individuals are grouped into strata --> M strata variable

  Strata or groups: regions, NACE sectors, professional categories, age groups ...

predictors

strata class

	sex - mult_n	sal25 - mult_n	b25 - mult_n	salh50 - mult_n	hm - mult_no	cvm - mult_no	on2 - nom	clerk - nc
4 e10	f (0.33), m (0.67)	YES (1.00)	YES (0.33), NO (0.67)	YES (1.00)	<=m (1.00)	<m (1.00)	elect	yes
4 e10	f (0.56), m (0.44)	YES (1.00)	YES (1.00)	YES (1.00)	<=m (1.00)	<m (1.00)	servi	yes
4 e10	f (0.50), m (0.50)	YES (1.00)	NO (1.00)	YES (1.00)	<=m (1.00)	<m (1.00)	servi	yes
4 e10	f (0.52), m (0.48)	YES (0.67), NO (0.33)	YES (0.14), NO (0.86)	YES (0.67), NO (0.33)	>m (0.67), <=m (0.33)	<m (1.00)	servi	yes

It represents a set of *clerk* employees of the sector NACE *services* who are:

52% *women*,

67% with *salary* lower than the first quartile

14% with month *bonuses* lower than the first quartile ... ...

Application Data

T25IT Italy: Monthly earnings by local unit size, NACE and ISCO - SES Data

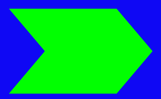
- Original Data:
  http://europa.eu.int/eu/comm/eurostat/research/conferences/ntts-98
  - 2772 segments representing 5.500.000 employees
  - M: NACE sector employee work for: Mining, manufacturing, electricity-gas-water, construction and services.
  - Z: clerk
  - $Y_1,...,Y_n$: sex, mean gross month earnings, idem per hour, mean weekly hours, mean gross month bonuses, size company....
- Data transformation
  - Binarisation of predictors
    - For each original predictor: 3 binary indicator variables for the first, second and third quartiles

Application Data

T25IT Italy: Monthly earnings by local unit size, NACE and ISCO - SES Data

- Two applications:
  - Nominal Data
  - Probabilistic data
    - Aggregation of segments by crossing categories of economic sector (22), profession (7), size company (7)
    - Selection of one predictor for each three related binary variables

- ==>720 data units represented by probabilistic data



(Both consider weights of original segments)

Input Data. Symbolic data

predictors ⟵ ⟶ strata class

	sex - mult_n	sal25 - mult_n	b25 - mult_n	salh50 - mult_n	hm - mult_no	cvm - mult_no	on2 - nom	clerk - nc
4 e10_	f (0.33), m (0.67)	YES (1.00)	YES (0.33), NO (0.67)	YES (1.00)	<=m (1.00)	<m (1.00)	elect	yes
4 e10_	f (0.56), m (0.44)	YES (1.00)	YES (1.00)	YES (1.00)	<=m (1.00)	<m (1.00)	servi	yes
4 e10_	f (0.50), m (0.50)	YES (1.00)	NO (1.00)	YES (1.00)	<=m (1.00)	<m (1.00)	servi	yes
4 e10_	f (0.52), m (0.48)	YES (0.67), NO (0.33)	YES (0.14), NO (0.86)	YES (0.67), NO (0.33)	=m (0.67), <=m (0.33)	<m (1.00)	servi	yes

→ It represents a set of *clerk* employees of the sector NACE *services* who are:

52% *women*,

67% with *salary* lower than the first quartile

14% with month *bonuses* lower than the first quartile ... ...

Objectives

- For the class variable:
  - Explain the classes by the predictors

    prediction rules (segmentation)

- For the strata:
  - Explain and classify the strata by common prediction rules

    Some rules are verified by some strata and not for others
  - Describe the strata by the prediction rules that can be applied in and by their weights

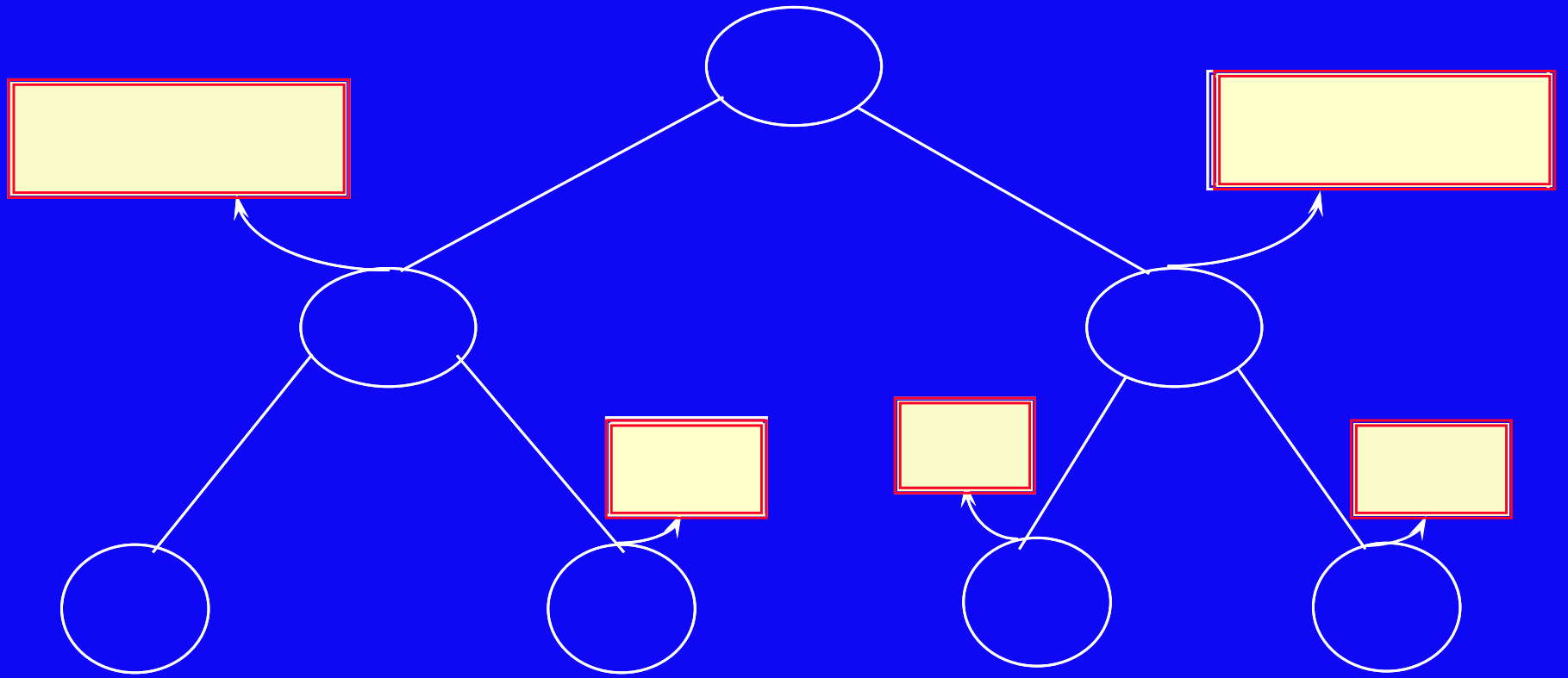
Application Data Objectives

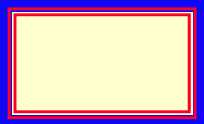
- Predict / explain profession clerk
- Obtain explanation of clerk employees by the predictors, conditioned by NACE sector (strata) they *belong to*
- Obtain sets of NACE sectors where this explanation is the same
- Describe a NACE sector by whole explanation of profession clerk

Strata decision tree

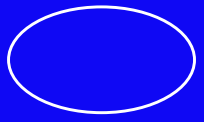
- Recursive tree building algorithm
- Consideration of strata in all steps
- Obtaining of class prediction rules, considering strata

Output. Tree





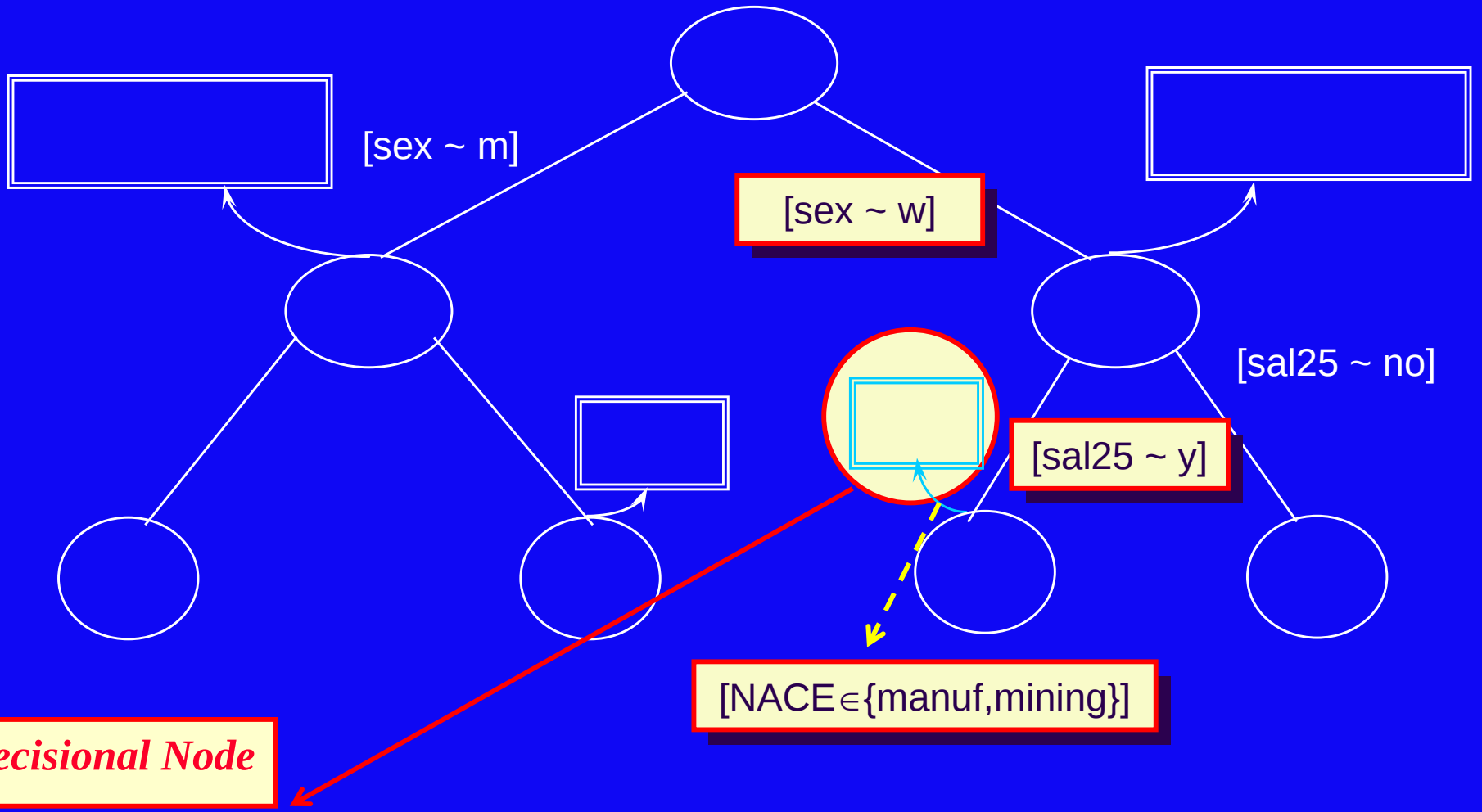
Decisional node: Terminal Node
+ Prediction rule



Exploratory node in some iteration

• • • • • • • • • • • •
• • • • • • • • •
• • • • • • • • • • • • •

Application Output. Tree



Decisional Node

predictors class strata

[sex ~ woman] ∧ [sal25 ~ yes] ∧[clerk ~(no (0.9), yes (0.1)] ∧[NACE ∈ {manuf,mining}]

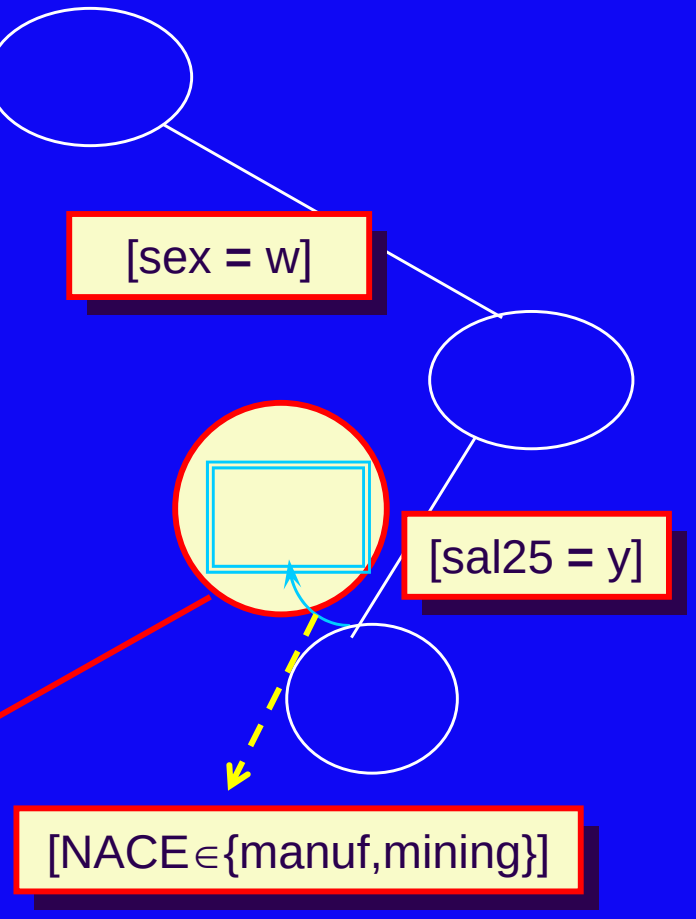
Application Output. Decisional Nodes

- Partition (monoevaluated data)

  sex(ω) = woman,

  sal25(ω) = yes,

  nace(ω) = manufacturing



predictors	strata
[sex = woman] ∧ [sal25 = yes]	∧[NACE ∈ {manuf,mining}]

Application Output. Decisional Nodes

- Partition with uncertainty (probabilistic data)

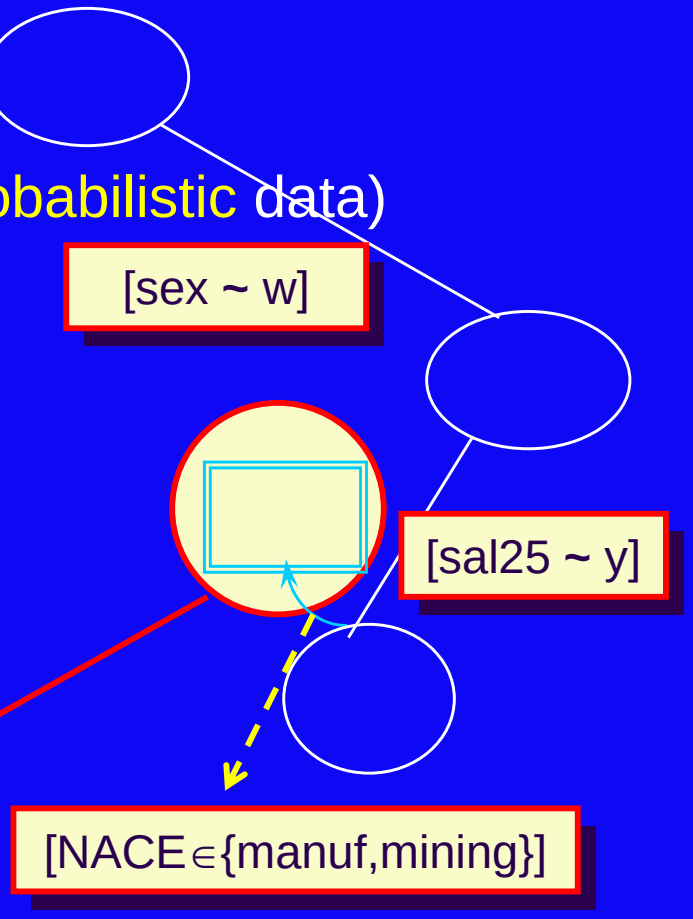
  sex(ω) = (woman(.8), man(.2)),

  sal25(ω) = (no(.2), yes(.8)),

  nace(ω) = manufacturing

Each node has a probability given ω

0.8x0.8x1=0.64



β_k - predictors

[sex ~ woman] ∧ [sal25 ~ yes]

μ_k - strata

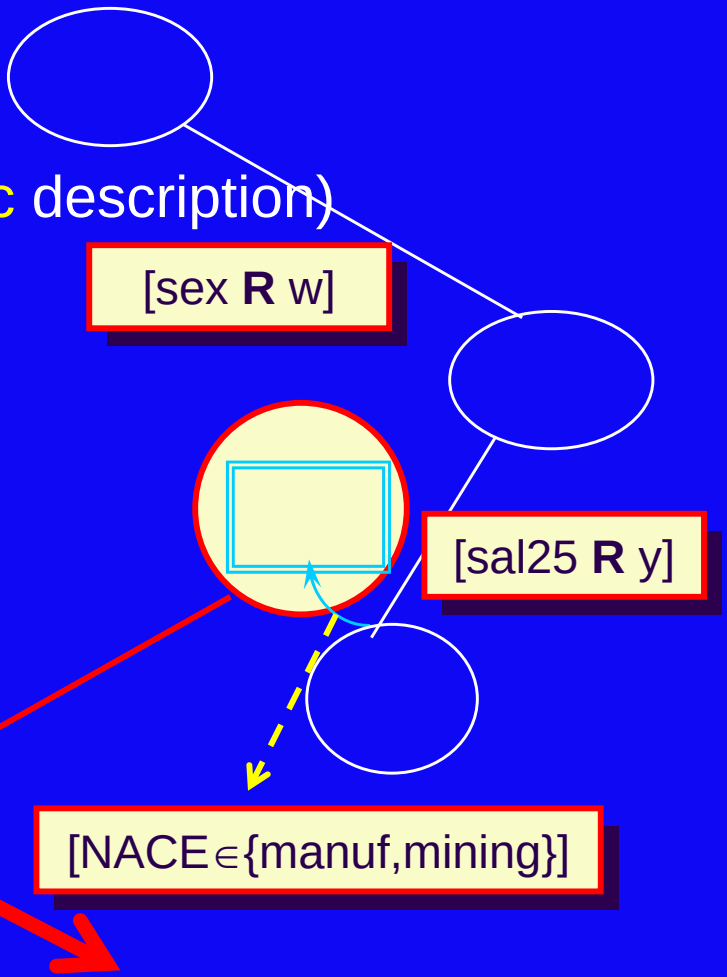
∧[NACE ∈ {manuf,mining}]

Application Output. Decisional Node

- Prediction for class (Symbolic description)
  - (no (0.9), yes (0.1))
  - Estimated probability for

    clerk is 0.1

    non-clerk is 0.9



predictors	class	strata
[sex **R** woman] ∧ [sal25 **R** yes]	[clerk ~(no (0.9), yes (0.1)]	∧[NACE ∈ {manuf,mining}]

Application Output. Decisional Node

- Prediction rule for a set of strata

For individuals in strata manufacturing and mining:

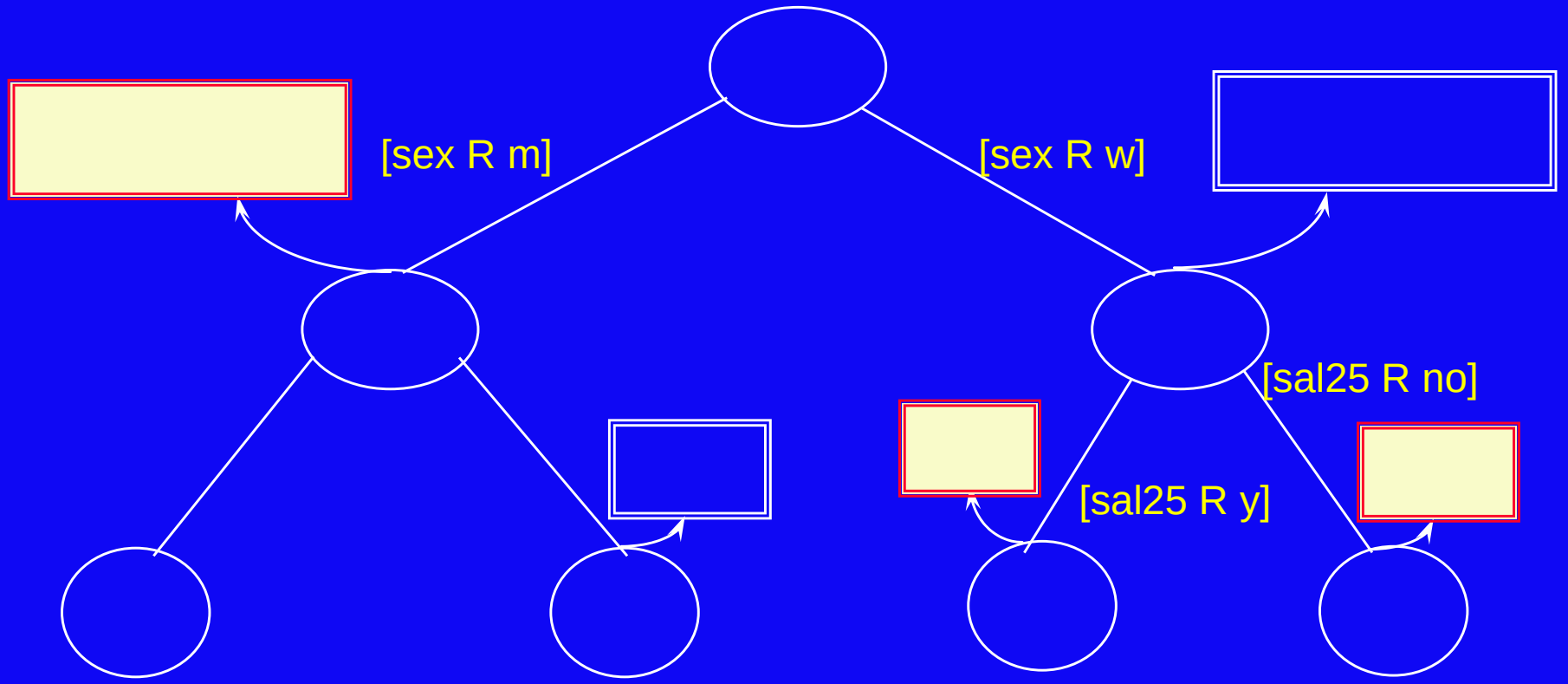
====> If [sex~woman] and [sal25 ~ yes]

====> Then estimated probability for clerk = no is 0.9

Decisional Node

predictors	class	strata
[sex ~ woman] $\wedge$ [sal25 ~ yes] $\wedge$[clerk ~(no (0.9), yes (0.1)] $\wedge$[NACE $\in$ {manuf,mining}]		

Application Output Strata Description



Manufacting: { 0.74 [sex R man] ∧ [clerk ~ (no(0.87),yes(0.13)],

0.17 [sex R woman] ∧ [salh25 R yes] ∧ [clerk ~ (no(0.9),yes(0.1)],

0.09 [sex R woman] ∧ [salh25 R no] ∧ [clerk ~ (no(0.26),yes(0.74)] }

Manufacturing stratum : 3 rules with respectives weights 0.74, 0.17 y 0.09

Application Output Strata Description

- For individuals in stratum manufacturing:

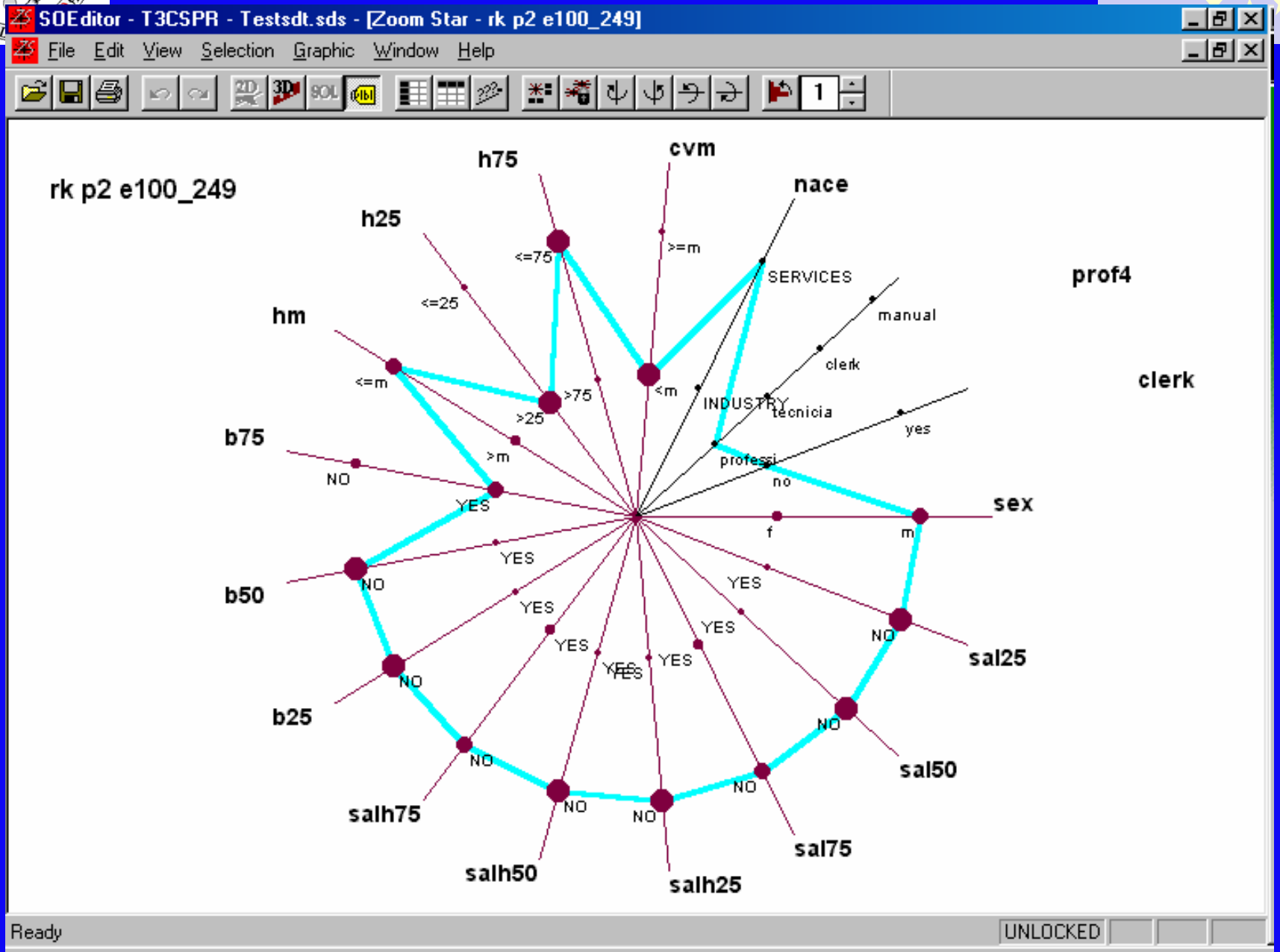
			estimated probability for		is	weight
If	sex R man	Then	clerk=no		0.87	0.74
If	sex R woman and salh25 R yes	Then	clerk=no		0.9	0.17
If	sex R woman and salh25 R no	Then	clerk=yes		0.74	0.09

Prediction

- Symbolic description
- Estimated probability for classes:
  - For nominal data:
    - Node class probabilities where individual *belongs to*
  - For probabilistic data:
    - Selection of decisional nodes with rules for the stratum where input data unit *belongs to*
    - Mixture of class probability distribution of nodes by probability of nodes given data unit
  - It is considered Missing values for predictors

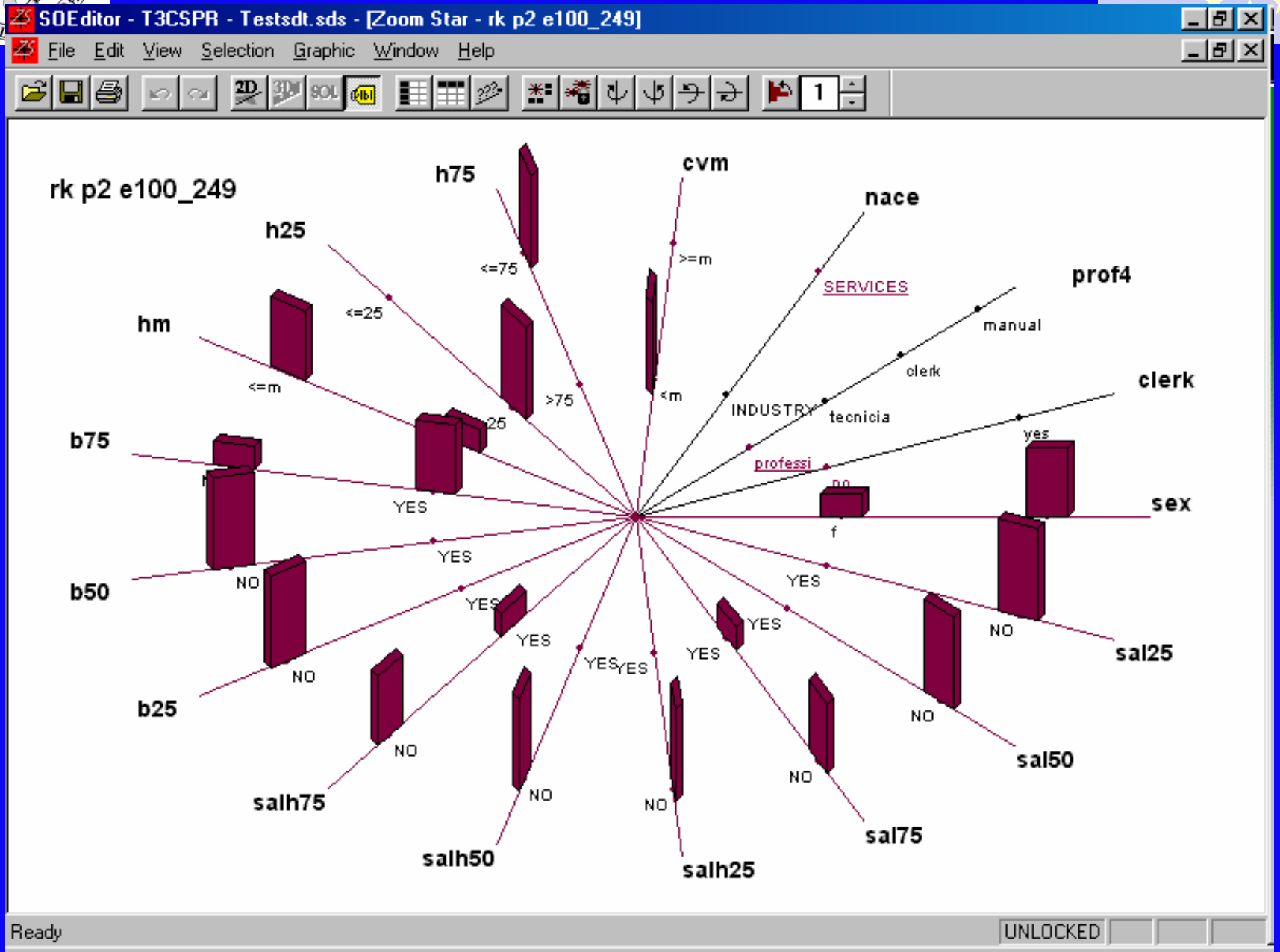
SOE (FUNDP)

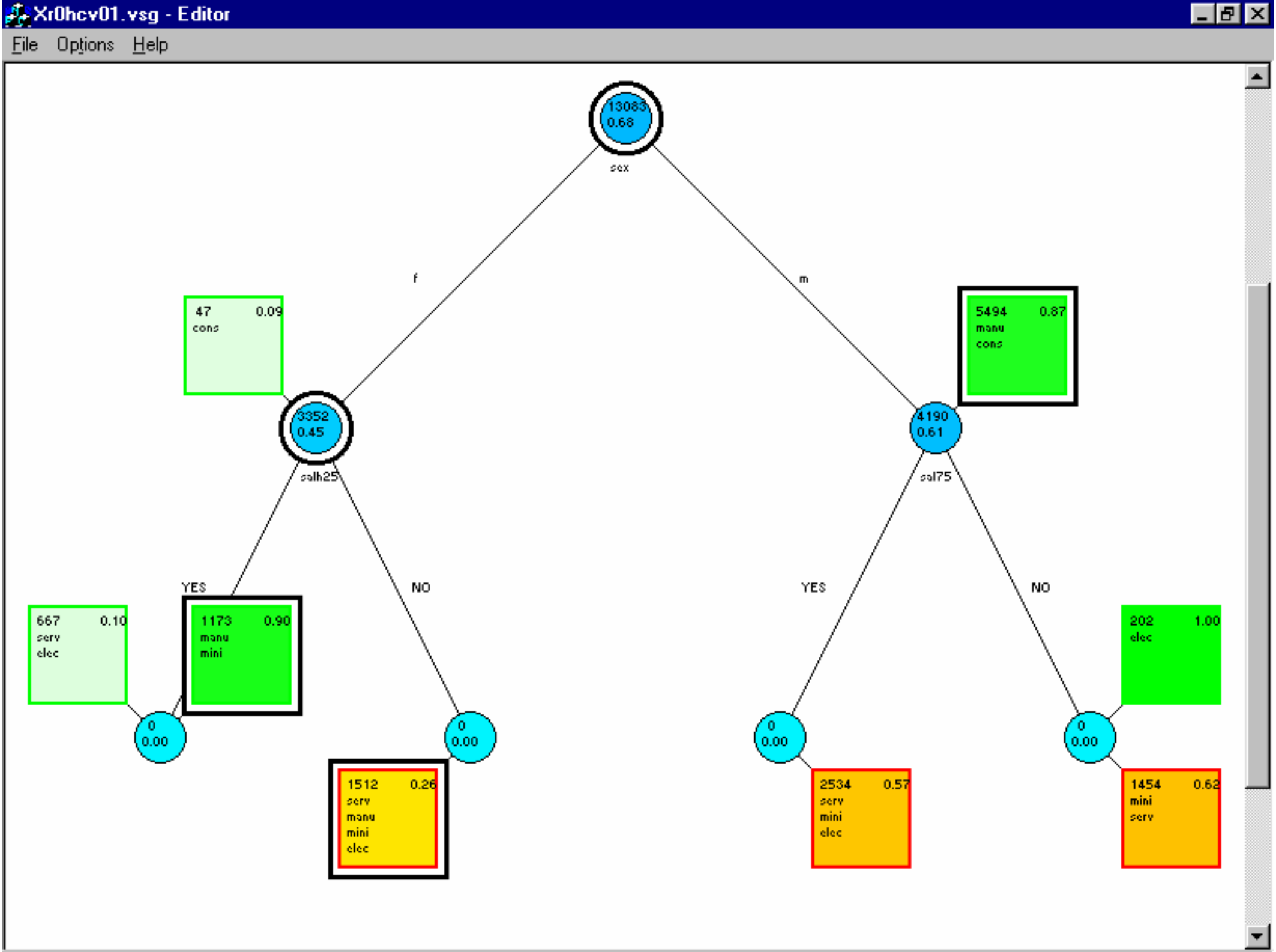
Input Symbolic Data Visualisation

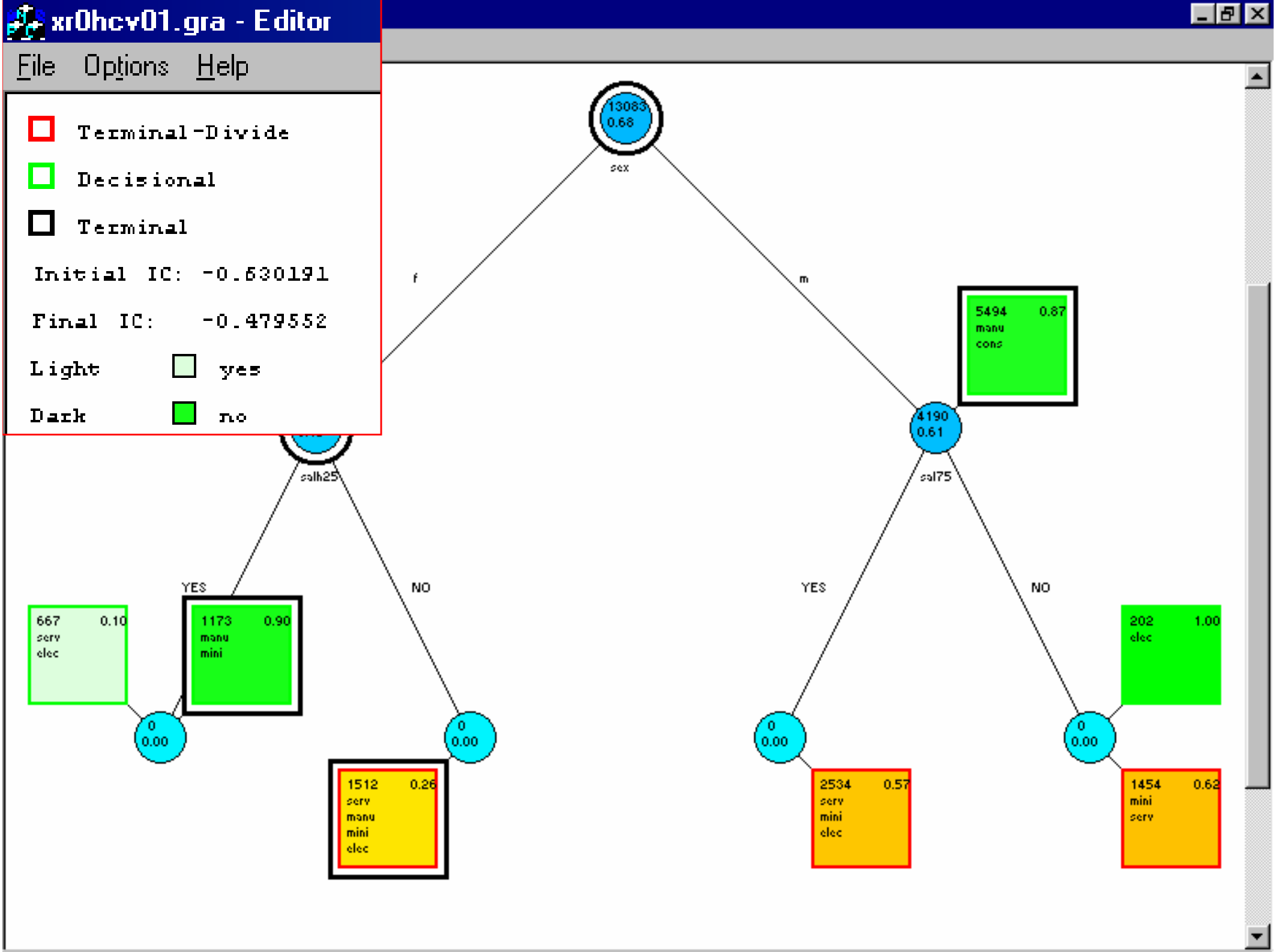


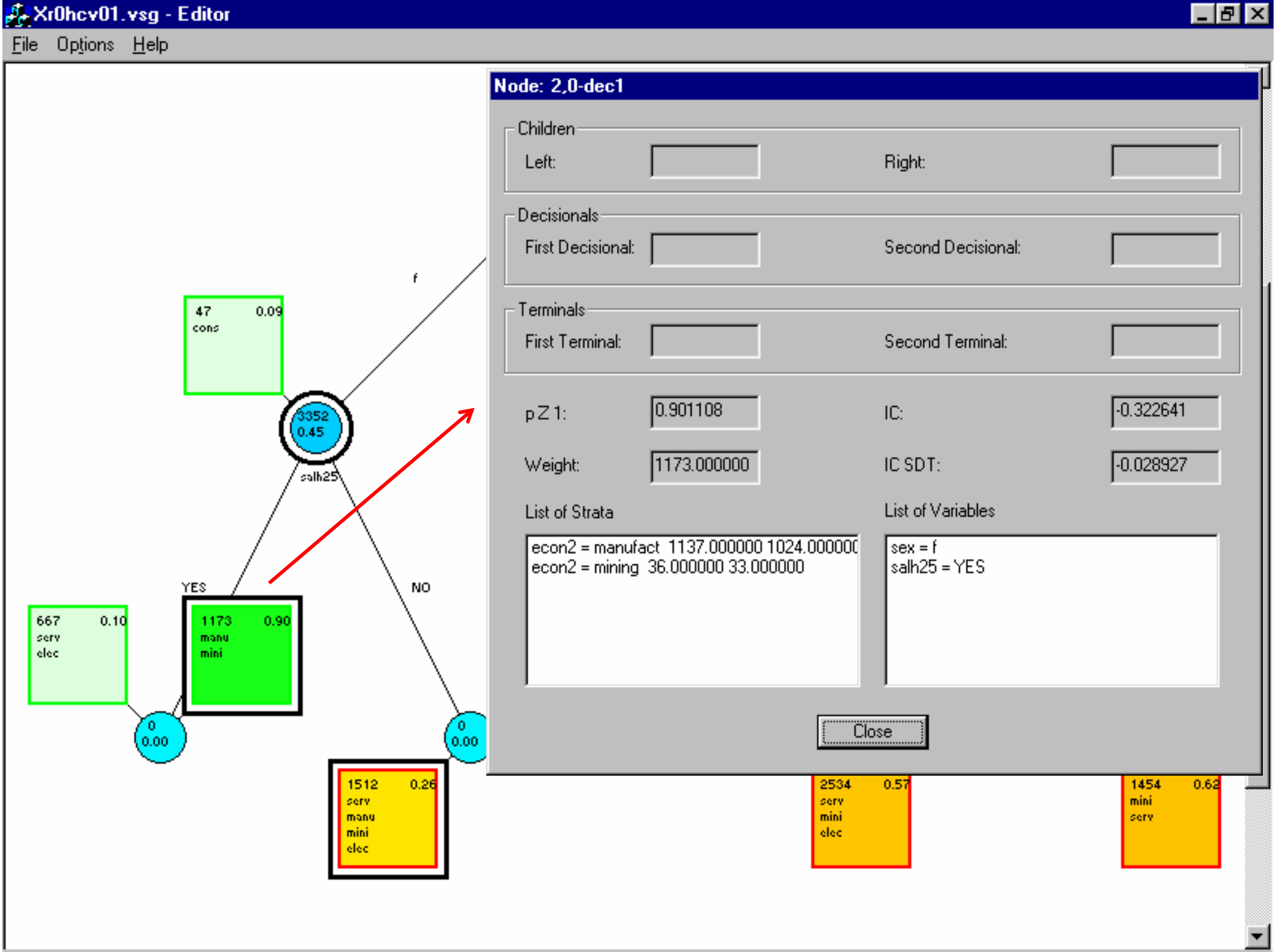
SOE (FUNDP)

Input Symbolic Data Visualisation




Xr0hcv01.vsg - Editor
File Options Help
13083
0.68
sex
f
m
47 0.09
cons
3352
0.45
salh25
4190
0.61
sal75
5494 0.87
manu
cons
YES
NO
YES
NO
667 0.10
serv
elec
1173 0.90
manu
mini
0
0.00
0
0.00
0
0.00
0
0.00
1512 0.26
serv
manu
mini
elec
2534 0.57
serv
mini
elec
202 1.00
elec
1454 0.62
mini
serv


xr0hcv01.gra - Editor
File Options Help
Terminal-Divide
Decisional
Terminal
Initial IC: -0.630191
Final IC: -0.479552
Light yes
Dark no
13083
0.68
sex
f
m
5494 0.87
manu
cons
4190
0.61
sal75
salh25
YES
NO
YES
NO
667 0.10
serv
elec
1173 0.90
manu
mini
0
0.00
0
0.00
1512 0.26
serv
manu
mini
elec
0
0.00
2534 0.57
serv
mini
elec
0
0.00
202 1.00
elec
1454 0.62
mini
serv


Xr0hcv01.vsg - Editor
File Options Help
Node: 2,0-dec1
Children
Left:
Right:
Decisionals
First Decisional:
Second Decisional:
Terminals
First Terminal:
Second Terminal:
p Z 1:
0.901108
IC:
-0.322641
Weight:
1173.000000
IC SDT:
-0.028927
List of Strata
List of Variables
econ2 = manufact 1137.000000 1024.000000
econ2 = mining 36.000000 33.000000
sex = f
salh25 = YES
Close
f
47 0.09
cons
3352
0.45
salh25
YES
NO
667 0.10
serv
elec
1173 0.90
manu
mini
0
0.00
0
0.00
1512 0.26
serv
manu
mini
elec
2534 0.57
serv
mini
elec
1454 0.62
mini
serv

Decisional nodes output

Decisional nodes (depth=2, p=0.83)

(ICInic = -0.63, ICFinal = -0.47)

report
sdteditor

10d0: [sex = f]^[NACE = {construc}]^[clerk ~ (0.09 no, 0.91 yes)]

11d1: [sex = m]^[NACE = {manufact, construc}]^[clerk ~ (0.87 no, 0.13 yes)]

20d0: [sex = f]^[salh25 = yes]^[NACE = {services, electric}]^[clerk ~ (0.10 no,0.90 yes)]

20d1: [sex = f]^[salh25 = yes]^[NACE = {manufact, mining}]^[clerk ~ (0.90 no, 0.10 yes)

23d1 : [sex = m]^[sal75 = no]^[NACE = {electric}]^[clerk ~ (no]

21td1 : [sex = f]^[salh25 = no]^[NACE = {services, manufact, mining, electric}]^
[clerk ~ (0.26 no, 0.74 yes)]

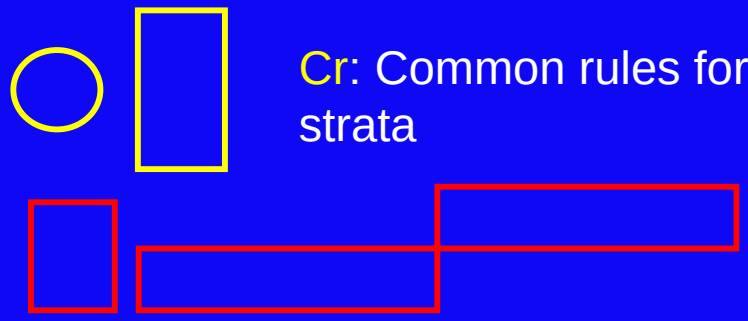
22td1: [sex = m]^[sal75 = y]^[NACE = {services, mining, electric}]^[clerk~(0.57 no, 0.43y)

23td1 : [sex = m]^[sal75 = no]^[NACE = {mining, services}]^[clerk ~ (0.62 no, 0.38 yes)]

Aids for interpretation

- Relative contributions (Cr) and absolute contributions (Ca)

	Minería		Manufactura		electricidad		construcción		servicios	
	Cr	Ca	Cr	Ca	Cr	Ca	Cr	Ca	Cr	Ca
10d0							0.08	1		
11d1			0.74	0.9			0.92	0.1		
20d0					x	0.01			0.13	0.99
20d1	0.16	0.04	0.17	0.96						
23d1					0.44	1				
21td0	0.14	0.02	0.09	0.39	0.11	0.03			0.17	0.56
22td1	0.56	0.05			0.45	0.08			0.43	0.87
23td1	0.14	0.02							0.27	0.98



Cr: Common rules for strata

Cr: Stratum characterisation

Ca: Importance of a stratum in a rule

Ca: Node characterisation by strata

20d0, 20d1: Anthagonic rules

Application on Free Time Consuming Survey

- Input:
  - 5.500 Individuals
  - Variables controlled by the survey: sex, age, day of week, region(3)
  - Other variables: relationship with job, profession, 32 time consuming aspects
- Several scenario analysed:
  - Class: sex
  - Strata: Profession | Day of Week (mo-th, fr, sa, su)
  - Predictors: 32 time consuming aspects

Application on Free Time Consuming Survey

- Aim:
  - Explain sex by their time consuming, conditioned by the profession/day of week
  - Identify sets of profession/day of week where sex explanation is the same
  - Describe a profession/day with the time consuming of both sex
- Other scenario analysed:
  - Class: Worker/non worker
  - Strata: day of week Predictors: time consuming aspects
  - Aim: the same, but to predict/explain one profession

Other application

- Perception of people about their own town
  - 20 towns (M), 4600 people
  - Global perception (Z)
  - Partial perceptions ($Y_1,\ldots,Y_n$)
- Labour activity
  - n = 5305 symbolic data units

Conclusion

- Input/Output in the framework of symbolic data analysis
- Management of groups of individuals (strata) in a tree-building algorithm that considers strata structure
- Strata description by prediction rules
- Classification/distintion of strata by common/different prediction rules
- More information in only one tree than in separate trees for each stratum :
  - Strata classification
  - Strata interpretation in the context of all strata
- Implementation of non-binary variables and other symbolic data: multivalued and interval data

References

- Bock, H.H., Diday E. (eds.) *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg
- Bravo, M.C. (2000) , Strata Decision Tree Symbolic Data Analysis Software, In: H.A.L. Kiers, J.P. Rasson, P.J.F Groenen, M. Shader (eds.) *Data Analysis, Classification and Related Methods*, Springer Verlag, 409-415
- Bravo Llatas, M.C. (2001), *Análisis de Segmentation en el Análisis de Datos Simbólicos*, Tesis doctoral, Universidad Complutense de Madrid.
- Bravo Llatas, M.C. García-Santesmases, J.M. (2000), Segmentation Trees for Stratified Data, In: H.H. Bock, E. Diday (eds.) , 266-293.
- ---------- (2000), Symbolic Object Description of Strata by Segmentation Trees, *Computational Statistics*, **15**, Physica-Verlag, 13-24
- Breiman, L., Friedman, J.H., Olshen, R.A. y Stone, C.J. (1984), *Classification and Regresion Trees*, Wadsworth International Group
- Périnel, E., Lechevallier, Y. (2000), Symbolic Discriminant Rules In: Bock, H.H., Diday, E. Eds., 2000, 244-265.