Learning the parameters of a multiple criteria sorting method from large sets of assignment examples

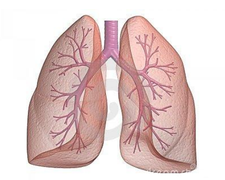
Olivier Sobrie[1,2] - Vincent Mousseau[1] - Marc Pirlot[2]

[1]École Centrale de Paris - Laboratoire de Génie Industriel
[2]University of Mons - Faculty of engineering

July 2, 2013

Introductory example

Application : Lung cancer



Categories :

C_3 : No cancer

C_2 : Curable cancer

C_1 : Incurable cancer

$$C_3 \succ C_2 \succ C_1$$

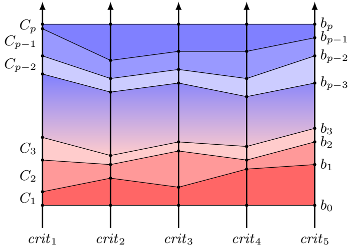
- 9394 patients analyzed
- Monotone attributes (number of cigarettes per day, age, ...)
- Output variable : no cancer, cancer, incurable cancer
- Predict the risk to get a lung cancer for other patients on basis of their attributes

MR-Sort procedure

Main characteristics

- Sorting procedure
- Simplified version of the ELECTRE TRI procedure [Yu, 1992]
- Axioms based [Slowínski et al., 2002, Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b]

Parameters



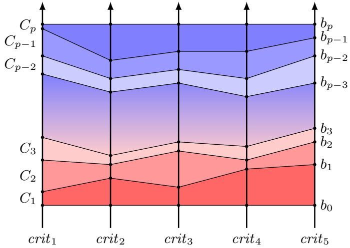
- Profiles' performances ($b_{h,j}$ for $h = 1, ..., p-1; j = 1, ..., n$)
- Criteria weights (w_j for $n = 1, ..., n$)
- Majority threshold (λ)

MR-Sort procedure

Main characteristics

- Sorting procedure
- Simplified version of the ELECTRE TRI procedure [Yu, 1992]
- Axioms based [Slowínski et al., 2002, Bouyssou and Marchant, 2007a, Bouyssou and Marchant, 2007b]

Parameters



Assignment rule

$$a \in C_h$$
$$\Leftrightarrow$$
$$\sum_{j:a_j \geq b_{h-1,j}} w_j \geq \lambda \text{ and } \sum_{j:a_j \geq b_{h,j}} w_j < \lambda$$

Inferring the parameters

What already exists to infer MR-Sort parameters ?

- Mixed Integer Program learning the parameters of an MR-Sort model [Leroy et al., 2011]
- Metaheuristic to learn the parameters of an ELECTRE TRI model [Doumpos et al., 2009]
- Not suitable for large problems : computing time becomes huge when the number of parameters or examples increases

Our objective

- Learn a MR-Sort model from a large set of assignment examples
- Efficient algorithm (i.e. can handle 1000 alternatives, 10 criteria, 5 categories)

Principe of the metaheuristic

Input parameters

- Assignment examples
- Performances of the examples on the *n* criteria

Objective

- Learn an MR-Sort model which is compatible with the highest number of assignment examples, i.e. maximize the classification accuracy,

$$CA = \frac{\text{Number of examples correctly restored}}{\text{Total number of examples}}$$

Difficulty

- Learn all the parameters of an MR-Sort model at the same time

Metaheuristic to learn all the parameters

Algorithm

Generate a population of N_{model} models with profiles initialized with a heuristic
repeat
  for all model M of the set **do**
    Learn the weights and majority threshold with a linear program, using the current profiles
    Adjust the profiles with a heuristic N_{it} times, using the current weights and threshold.
  end for
  Reinitialize the $\left\lfloor \frac{N_{model}}{2} \right\rfloor$ models giving the worst CA
until Stopping criterion is met

Stopping criterion

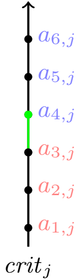
Stopping criterion is met when one model has a CA equal to 1 or when the algorithm has run N_o times.

Profiles initialization

Principe

- By a heuristic
- On each criterion j, give to the profile a performance such that CA would be max for the alternatives belonging to h and $h+1$ if $w_j = 1$.
- Take the probability to belong to a category into account

Example 1 : Where should the profile be set on criterion j ?



$a_{i,j}$	Category
$a_{1,j}$	C_1
$a_{2,j}$	C_1
$a_{3,j}$	C_1
$a_{4,j}$	C_2
$a_{5,j}$	C_2
$a_{6,j}$	C_2

Category	$P(a_i \in C_h)$
C_1	$\frac{1}{2}$
C_2	$\frac{1}{2}$

$C_2 \succ C_1$

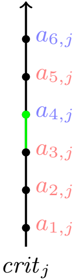
$$a_{3,j} < b_h \leq a_{4,j}$$

Profiles initialization

Principe

- By a heuristic
- On each criterion j, give to the profile a performance such that CA would be max for the alternatives belonging to h and $h+1$ if $w_j = 1$.
- Take the probability to belong to a category into account

Example 2 : Where should the profile be set on criterion j ?



$a_{i,j}$	Category
$a_{1,j}$	C_1
$a_{2,j}$	C_1
$a_{3,j}$	C_1
$a_{4,j}$	C_2
$a_{5,j}$	C_1
$a_{6,j}$	C_2

Category	$P(a_i \in C_h)$
C_1	$\frac{2}{3}$
C_2	$\frac{1}{3}$

$$C_2 \succ C_1$$

$$\boxed{a_{3,j} < b_h \leq a_{4,j}}$$

Learning the weights and the majority threshold

Principe

- Maximizing the classification accuracy of the model
- Using a linear program with no binary variables

Linear program

$$\text{Objective} : \min \sum_{a_i \in A} (x'_i + y'_i) \qquad (1)$$

$$\sum_{\forall j | a_i S_j b_{h-1}} w_j - x_i + x'_i = \lambda \qquad \forall a_i \in A_h, h = \{2, ..., p-1\} \qquad (2)$$

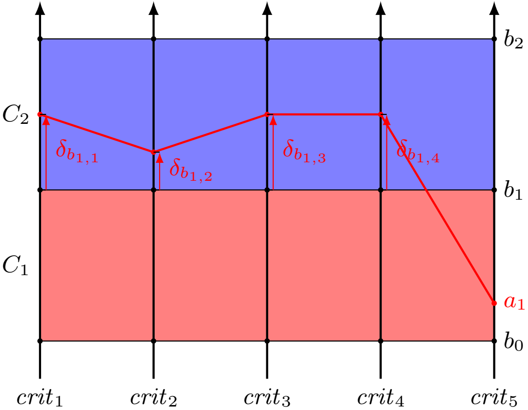
$$\sum_{\forall j | a_i S_j b_h} w_j + y_i - y'_i = \lambda - \delta \qquad \forall a_i \in A_h, h = \{1, ..., p-2\} \qquad (3)$$

$$\sum_{j=1}^{n} w_j = 1 \qquad (4)$$

(5)

Learning the profiles

Case 1 : Alternative a_1 classified in C_2 instead of C_1 ($C_2 \succ C_1$)

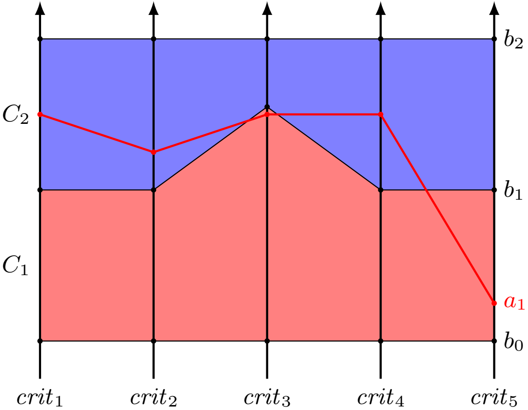


- a_1 is classified by the **DM** into category C_1
- a_1 is classified by the **model** into category C_2
- a_1 outranks b_1
- Profile too low on one or several criteria (in red)

$w_j = 0.2$ for $j = 1, ..., 5$; $\lambda = 0.8$

Learning the profiles

Case 1 : Alternative a_1 classified in C_2 instead of C_1 ($C_2 \succ C_1$)

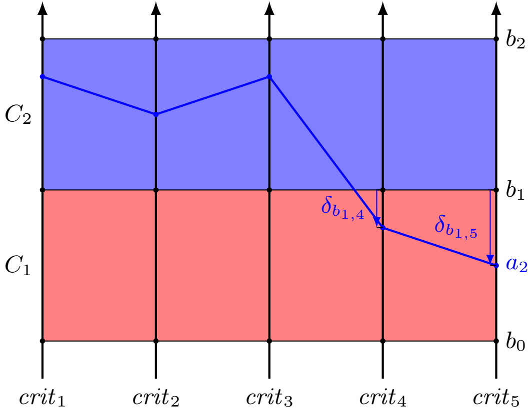


- a_1 is classified by the **DM** into category C_1
- a_1 is classified by the **model** into category C_2
- a_1 outranks b_1
- Profile too low on one or several criteria (in red)

$w_j = 0.2$ for $j = 1, ..., 5$; $\lambda = 0.8$

Learning the profiles

Case 2 : Alternative a_2 classified in C_1 instead of C_2 ($C_2 \succ C_1$)

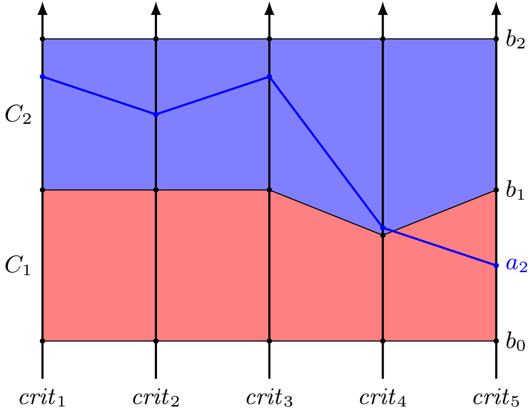


$w_j = 0.2$ for $j = 1, \ldots, 5$; $\lambda = 0.8$

- a_2 is classified by the **DM** into category C_2
- a_2 is classified by the **model** into category C_1
- a_2 doesn't outrank b_1
- Profile too high on one or several criteria (in blue)
- If profile moved by $\delta_{b_1,2,4}$ on g_4 and/or by $\delta_{b_1,2,5}$ on g_5, the alternative will be rightly classified

Learning the profiles

Case 2 : Alternative a_2 classified in C_1 instead of C_2 ($C_2 \succ C_1$)

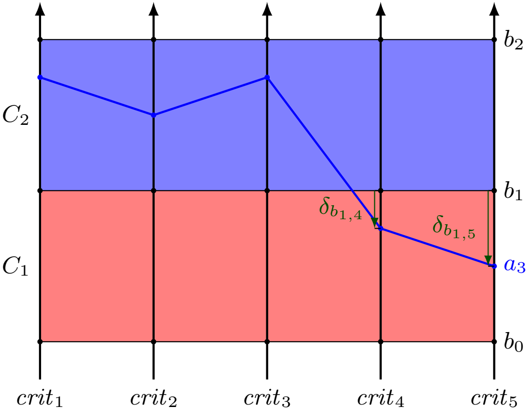


$w_j = 0.2$ for $j = 1, ..., 5$; $\lambda = 0.8$

- a_2 is classified by the **DM** into category C_2
- a_2 is classified by the **model** into category C_1
- a_2 doesn't outrank b_1
- Profile too high on one or several criteria (in blue)
- If profile moved by $\delta_{b_1,2,4}$ on g_4 and/or by $\delta_{b_1,2,5}$ on g_5, the alternative will be rightly classified

Learning the profiles

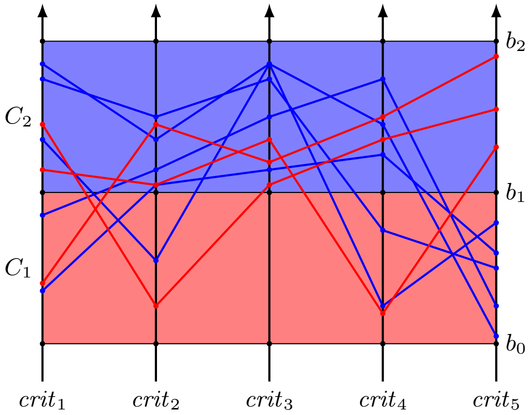
- $V_{h,j}^{+\delta}$ (resp. $V_{h,j}^{-\delta}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h by $+\delta$ (resp. $-\delta$) on j results in a correct assignment.



- $C_2 \succ C_1$
- $w_j = 0.2$ for $j = 1, ..., 5$
- $\lambda = 0.8$
- $a_3 \in A_{2 \leftarrow \text{DM}}^{1 \leftarrow \text{Model}}$

Learning the profiles

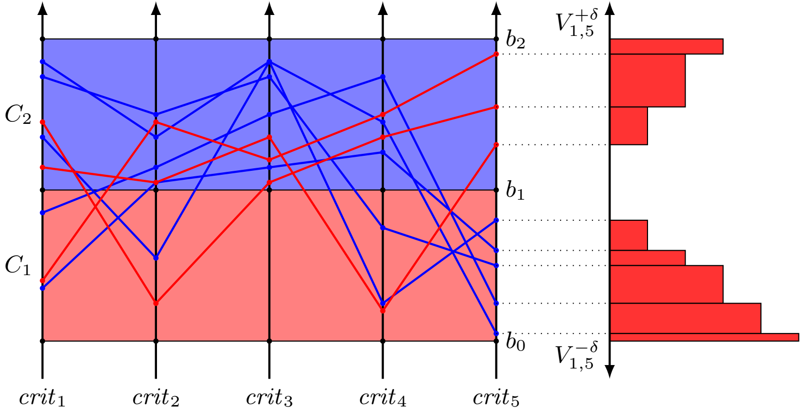
- $V_{h,j}^{+\delta}$ (resp. $V_{h,j}^{-\delta}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h by $+\delta$ (resp. $-\delta$) on j results in a correct assignment.



- $C_2 \succ C_1$
- $w_j = 0.2$ for $j = 1, ..., 5$
- $\lambda = 0.8$
- $a_3 \in A_{2 \leftarrow \text{DM}}^{1 \leftarrow \text{Model}}$

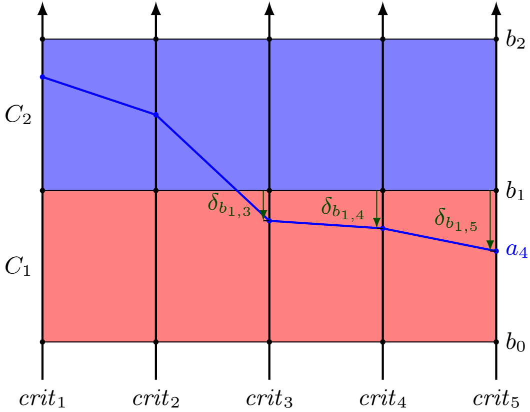
Learning the profiles

- $V_{h,j}^{+\delta}$ (resp. $V_{h,j}^{-\delta}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h by $+\delta$ (resp. $-\delta$) on j results in a correct assignment.



Learning the profiles

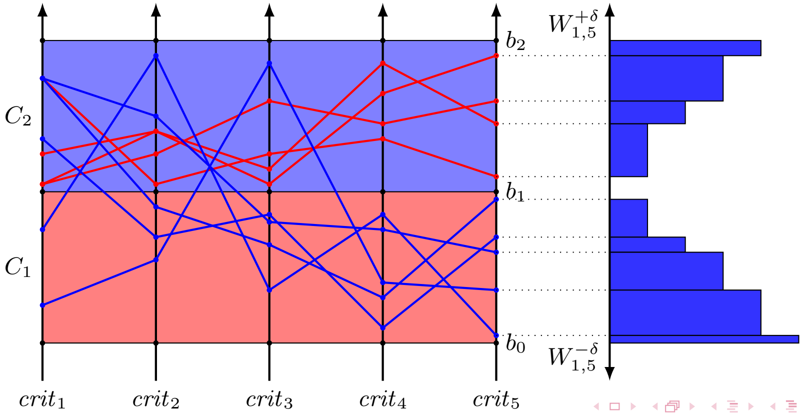
- $W_{h,j}^{+\delta}$ (resp. $W_{h,j}^{-\delta}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j strengthens the criteria coalition in favor of the correct classification but will not by itself result in a correct assignment.



- $C_2 \succ C_1$
- $w_j = 0.2$ for $j = 1, ..., 5$
- $\lambda = 0.8$
- $a_4 \in A_{2 \leftarrow \text{DM}}^{1 \leftarrow \text{Model}}$

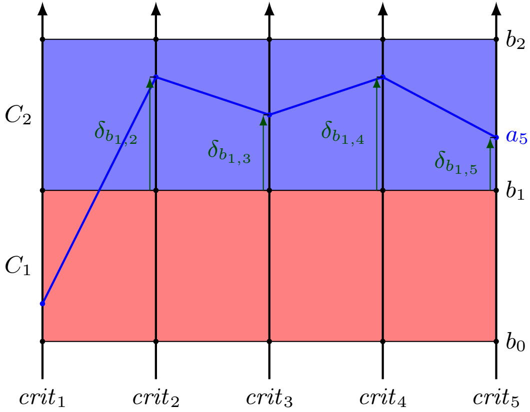
Learning the profiles

- $W_{h,j}^{+\delta}$ (resp. $W_{h,j}^{-\delta}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j strengthens the criteria coalition in favor of the correct classification but will not by itself result in a correct assignment.



Learning the profiles

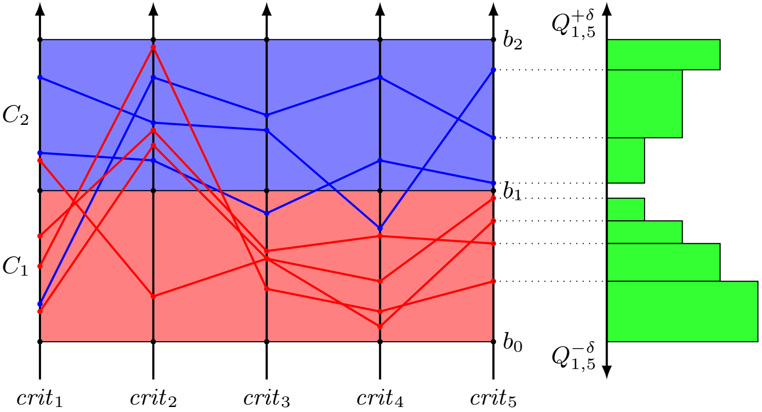
- $Q_{h,j}^{+\delta}$ (resp. $Q_{h,j}^{-\delta}$) : the sets of alternatives correctly classified in C_{h+1} (resp. C_{h+1}) for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j results in a misclassification.



- $C_2 \succ C_1$
- $w_j = 0.2$ for $j = 1, ..., 5$
- $\lambda = 0.8$
- $a_5 \in A_{2 \leftarrow DM}^{2 \leftarrow Model}$

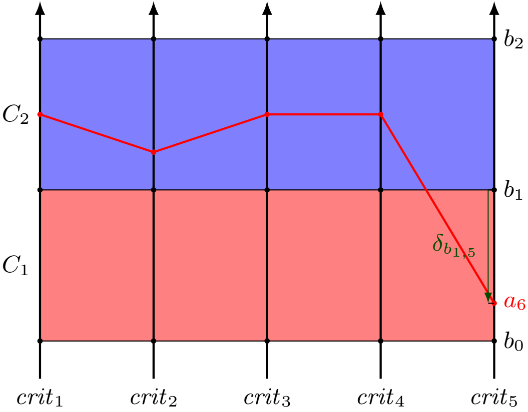
Learning the profiles

- $Q^{+\delta}_{h,j}$ (resp. $Q^{-\delta}_{h,j}$) : the sets of alternatives correctly classified in C_{h+1} (resp. C_{h+1}) for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j results in a misclassification.



Learning the profiles

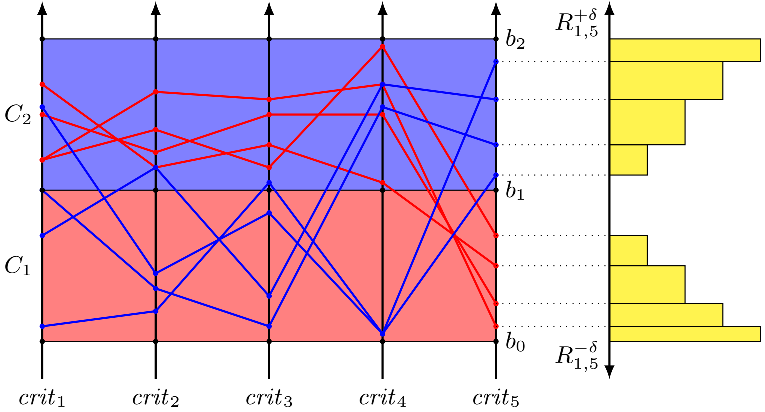
- $R^{+\delta}_{h,j}$ (resp. $R^{-\delta}_{h,j}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j weakens the criteria coalition in favor of the correct classification but does not induce misclassification by itself.



- $C_2 \succ C_1$
- $w_j = 0.2$ for $j = 1, ..., 5$
- $\lambda = 0.8$
- $a_6 \in A^{1 \leftarrow \text{Model}}_{2 \leftarrow \text{DM}}$

Learning the profiles

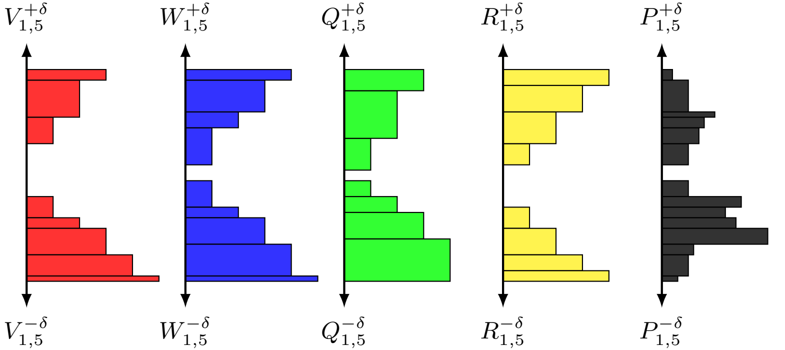
- $R^{+\delta}_{h,j}$ (resp. $R^{-\delta}_{h,j}$) : the sets of alternatives misclassified in C_{h+1} instead of C_h (resp. C_h instead of C_{h+1}), for which moving the profile b_h of $+\delta$ (resp. $-\delta$) on j weakens the criteria coalition in favor of the correct classification but does not induce misclassification by itself.



Learning the profiles

$$P(b_{1,j}^{+\delta}) = \frac{k_V|V_{1,j}^{+\delta}| + k_W|W_{1,j}^{+\delta}| + k_T|T_{1,j}^{+\delta}|}{d_V|V_{1,j}^{+\delta}| + d_W|W_{1,j}^{+\delta}| + d_T|T_{1,j}^{+\delta}| + d_Q|Q_{1,j}^{+\delta}| + d_R|R_{1,j}^{+\delta}|}$$

with : $k_V = 2, k_W = 1, k_T = 0.1, d_V = d_W = d_T = 1, d_Q = 5, d_R = 1$



Learning the profiles

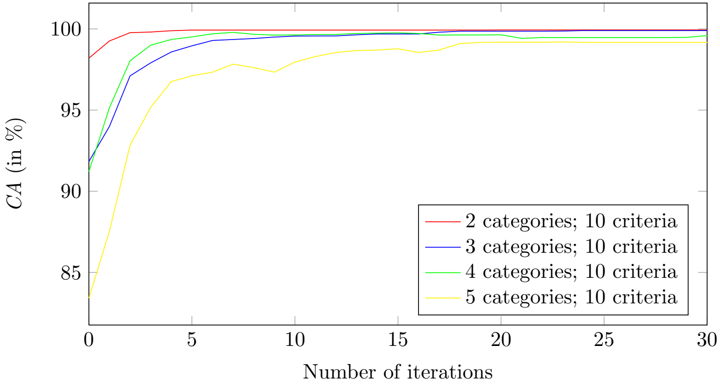
Overview of the complete algorithm

for all profile b_h **do**
  for all criterion j chosen randomly **do**
    Choose, in a randomized manner, a set of positions in the interval $[b_{h-1,j}, b_{h+1,j}]$
    Select the one such that $P(b_{h,j}^{\Delta})$ is maximal
    Draw uniformly a random number r from the interval $[0, 1]$.
    if $r \leq P(b_{h,j}^{\Delta})$ **then**
      Move $b_{h,j}$ to the position corresponding to $b_{h,j} + \Delta$
      Update the alternatives assignment
    end if
  end for
end for

Experimentations

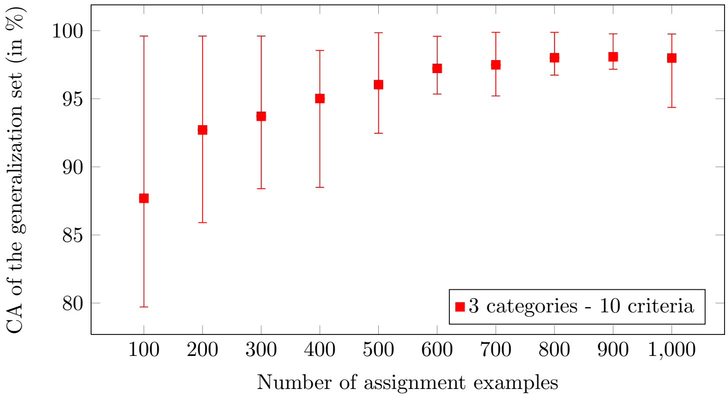
1. What's the efficiency of the algorithm ?
2. How much alternatives are required to learn a good model ?
3. What's the capability of the algorithm to restore assignments when there are errors in the examples ?
4. How the algorithm behaves on real datasets ?

Algorithm efficiency



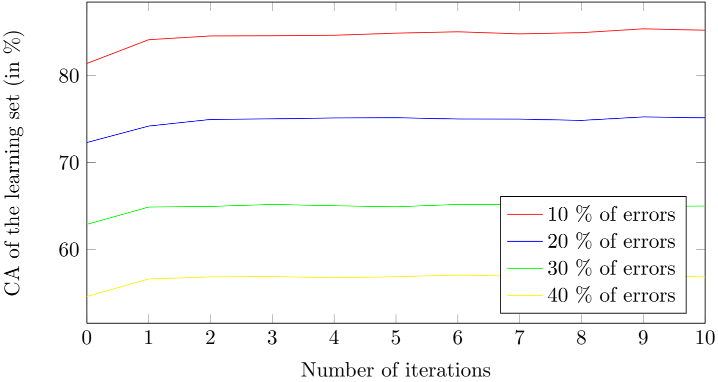
- Random model M generated
- Learning set : random alternatives assigned through the model M
- Model M' learned with the metaheuristic from the learning set

Model retrieval



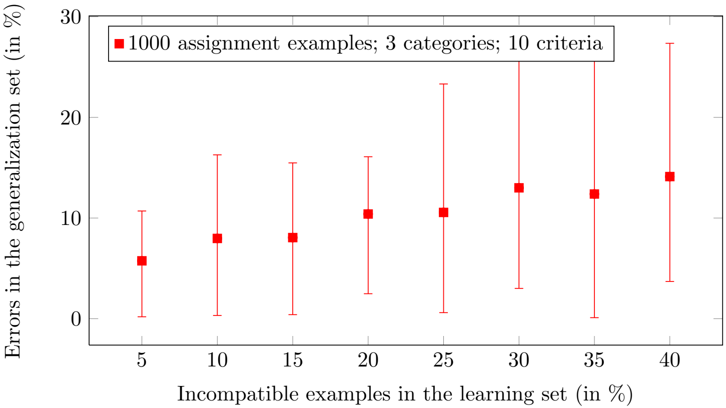
- Random model M generated
- Learning set : random alternatives assigned through model M
- Model M' learned with the metaheuristic from the learning set
- Generalization set : random alternatives assigned through M and M'

Tolerance for errors



- Random model M generated
- Learning set : random alternatives assigned through model M + errors
- Model M' learned with the metaheuristic from the learning set

Tolerance for errors



- Random model M generated
- Learning set : random alternatives assigned through model M + errors
- Model M' learned with the metaheuristic from the learning set
- Generalization set : random alternatives assigned through M and M'

Application on real datasets

Dataset	#instances	#attributes	#categories
DBS	120	8	2
CPU	209	6	4
BCC	286	7	2
MPG	392	7	36
ESL	488	4	9
MMG	961	5	2
ERA	1000	4	4
LEV	1000	4	5
CEV	1728	6	4

- Instances split in two parts : learning and generalization sets
- Binarization of the categories

Source : [Tehrani et al., 2012]

Application on real datasets - Binarized categories

Learning set	Dataset	MIP MR-SORT	META MR-SORT	LP UTADIS
20 %	DBS	0.8023 ± 0.0481	0.8012 ± 0.0469	0.7992 ± 0.0533
	CPU	0.9100 ± 0.0345	0.8960 ± 0.0433	0.9348 ± 0.0362
	BCC	0.7322 ± 0.0276	0.7196 ± 0.0302	0.7085 ± 0.0307
	MPG	0.7920 ± 0.0326	0.7855 ± 0.0383	0.7775 ± 0.0318
	ESL	0.8925 ± 0.0158	0.8932 ± 0.0159	0.9111 ± 0.0160
	MMG	0.8284 ± 0.0140	0.8235 ± 0.0135	0.8160 ± 0.0184
	ERA	0.7907 ± 0.0174	0.7915 ± 0.0146	0.7632 ± 0.0187
	LEV	0.8386 ± 0.0151	0.8327 ± 0.0221	0.8346 ± 0.0160
	CEV	-	0.9214 ± 0.0045	0.9206 ± 0.0059
50 %	DBS	0.8373 ± 0.0426	0.8398 ± 0.0487	0.8520 ± 0.0421
	CPU	0.9360 ± 0.0239	0.9269 ± 0.0311	0.9770 ± 0.0238
	BCC	-	0.7246 ± 0.0446	0.7146 ± 0.0246
	MPG	-	0.8170 ± 0.0295	0.7910 ± 0.0236
	ESL	0.8982 ± 0.0155	0.8982 ± 0.0203	0.9217 ± 0.0163
	MMG	-	0.8290 ± 0.0153	0.8242 ± 0.0152
	ERA	0.8042 ± 0.0137	0.7951 ± 0.0191	0.7658 ± 0.0171
	LEV	0.8554 ± 0.0151	0.8460 ± 0.0221	0.8444 ± 0.0132
	CEV	-	0.9216 ± 0.0067	0.9201 ± 0.0091
80 %	DBS	0.8520 ± 0.0811	0.8712 ± 0.0692	0.8720 ± 0.0501
	CPU	0.9402 ± 0.0315	0.9476 ± 0.0363	0.9848 ± 0.0214
	BCC	-	0.7486 ± 0.0640	0.7087 ± 0.0510
	MPG	-	0.8152 ± 0.0540	0.7920 ± 0.0388
	ESL	0.8992 ± 0.0247	0.9017 ± 0.0276	0.9256 ± 0.0235
	MMG	-	0.8313 ± 0.0271	0.8266 ± 0.0265
	ERA	0.8144 ± 0.0260	0.7970 ± 0.0272	0.7644 ± 0.0292
	LEV	0.8628 ± 0.0232	0.8401 ± 0.0321	0.8428 ± 0.0222
	CEV	-	0.9204 ± 0.0130	0.9201 ± 0.0132

Application on real datasets

	Dataset	MIP MR-SORT	META MR-SORT	LP UTADIS
20 %	CPU	0.7542 ± 0.0506	0.7443 ± 0.0559	0.8679 ± 0.0488
	ERA	-	0.5104 ± 0.0198	0.4856 ± 0.0169
	LEV	-	0.5528 ± 0.0274	0.5775 ± 0.0175
	CEV	-	0.7761 ± 0.0183	0.7719 ± 0.0153
50 %	CPU	-	0.8052 ± 0.0361	0.9340 ± 0.0266
	ERA	-	0.5216 ± 0.0180	0.4833 ± 0.0171
	LEV	-	0.5751 ± 0.0230	0.5889 ± 0.0158
	CEV	-	0.7833 ± 0.0180	0.7714 ± 0.0158
80 %	CPU	-	0.8055 ± 0.0560	0.9512 ± 0.0351
	ERA	-	0.5230 ± 0.0335	0.4824 ± 0.0332
	LEV	-	0.5750 ± 0.0344	0.5933 ± 0.0305
	CEV	-	0.7895 ± 0.0203	0.7717 ± 0.0259

Conclusions and further research

- Algorithm able to handle large datasets
- Adapted to the structure of the problem

- Comparison of AVF-Sort and MR-Sort
- Use MR-Sort models with vetoes
- Test the algorithm on other datasets


That's all Folks!

References I

Bouyssou, D. and Marchant, T. (2007a).
An axiomatic approach to noncompensatory sorting methods in MCDM, I : The case of two categories.
European Journal of Operational Research, 178(1) :217–245.

Bouyssou, D. and Marchant, T. (2007b).
An axiomatic approach to noncompensatory sorting methods in MCDM, II : More than two categories.
European Journal of Operational Research, 178(1) :246–276.

Doumpos, M., Marinakis, Y., Marinaki, M., and Zopounidis, C. (2009).
An evolutionary approach to construction of outranking models for multicriteria classification : The case of the ELECTRE TRI method.
European Journal of Operational Research, 199(2) :496–505.

References II

Leroy, A., Mousseau, V., and Pirlot, M. (2011).
Learning the parameters of a multiple criteria sorting method.
In Brafman, R., Roberts, F., and Tsoukiàs, A., editors, *Algorithmic Decision Theory*, volume 6992 of *Lecture Notes in Computer Science*, pages 219–233. Springer Berlin / Heidelberg.

Slowínski, R., Greco, S., and Matarazzo, B. (2002).
Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle.
Control and Cybernetics, 31(4) :1005–1035.

Tehrani, A. F., Cheng, W., Dembczynski, K., and Hüllermeier, E. (2012).
Learning monotone nonlinear models using the choquet integral.
Machine Learning, 89(1-2) :183–211.

References III

Yu, W. (1992).
Aide multicritère à la décision dans le cadre de la problématique du tri : méthodes et applications.
PhD thesis, LAMSADE, Université Paris Dauphine, Paris.