

ICML 2004, PDMC Challenge

ROC-based Genetic Learning

Jérôme Azé, Noël Lucas, Michèle Sebag

LRI, Université Paris-Sud

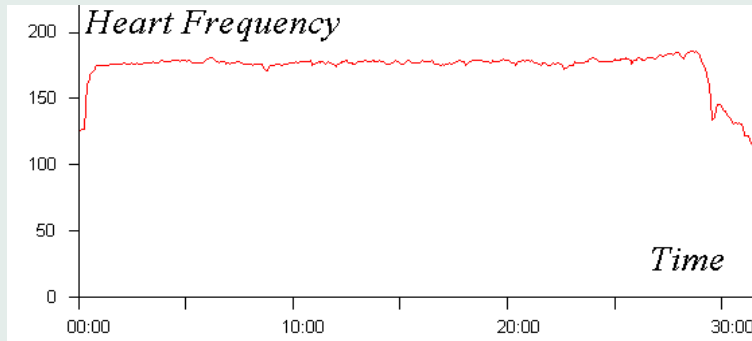
Working hypotheses

Physiological considerations:

sensor values correlated with the activity

same for heart beat

good stability along time



Cooper et al., 1985

⇒ Take the average sensor value

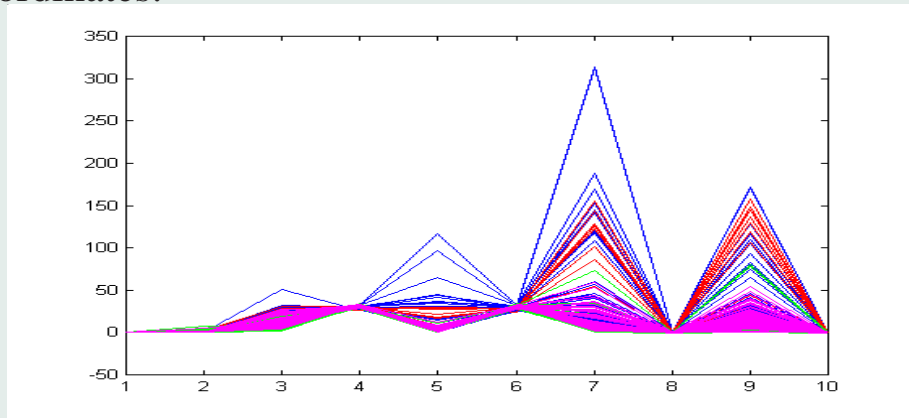
Outline

- Data preparation
- Algorithm used: ROGER
ROC-based Evolutionary learner
- Results

Data preparation, 1

Visualizing the data

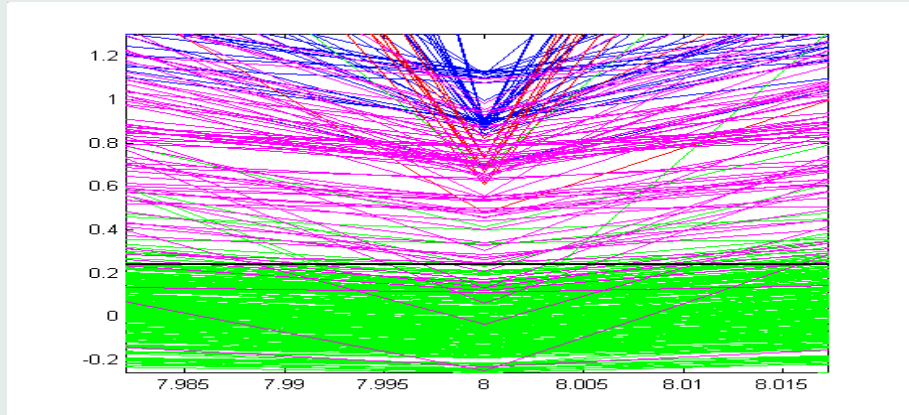
Parallel coordinates:



Visual rules

Some activities → simple profile

rules can be extracted by experts : *Activity 3004 if $S_7 > .225$*

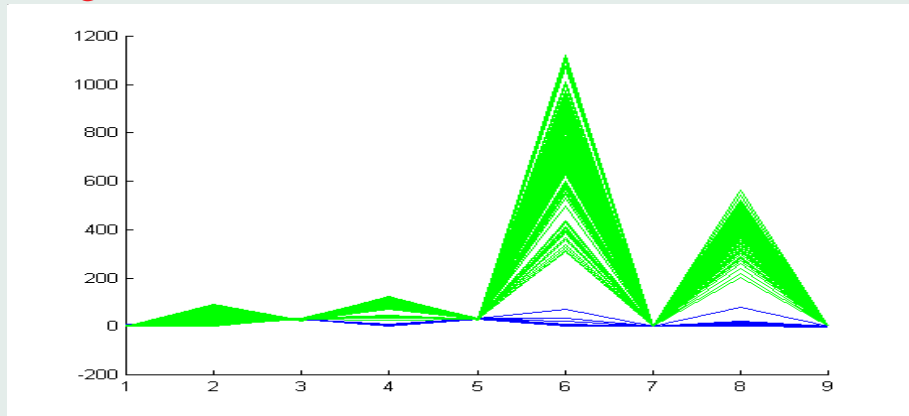


But mostly univariate analysis

very limited

More Visual Analysis

Activity 1201 against 5102



Data preparation, 2

Cleaning the data : remove cases where

$sensor_5 = -273$, or

$individual = 9$ and $activity = 3101$

Not using unlabeled data

remove $activity = 0$

(reduces computational time from days to hours)

Normalization

each sensor ranges in $[0,1]$

Constructing training sets

Learning Gender

Each session, activity \rightarrow an example

Each sensor \rightarrow its average value

Learning Activities

binary concept learning

Each activity $A \rightarrow$ training set \mathcal{E}_A

for each session, remove data from first and last minutes

each three minutes window \rightarrow an example

each sensor \rightarrow its average value.

label = (activity performed = A)

ROGER

Given $\mathcal{E} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}, i = 1 \dots n\}$

Learning criterion

The area under the ROC curve \equiv Wilcoxon statistics

$$\mathcal{F}(h) = Pr(h(\mathbf{x}_i) > h(\mathbf{x}_j) \mid y_i > y_j)$$

Search space : linear hypotheses

$$h = \mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d, \quad h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$$

References

Artificial Evolution VI, 2003

IEEE-ICDM, 2003

Bagged non-linear ROGER

Search space : non-linear hypotheses

$$h = (w_1, \dots, w_d, c_1, \dots, c_d) \in \mathbb{R}^{2d}, \quad h(\mathbf{x}) = \sum_{i=1}^d w_i |x_i - c_i|$$

Ensemble learning w. artificial evolution

each run \rightarrow hypothesis h

Given h_1, \dots, h_T , define

$$BH(\mathbf{x}) = \text{Median}\{h_1(\mathbf{x}), \dots, h_T(\mathbf{x})\}$$

Experimental setting

10 fold cross-validation

all examples derived from a session in the same fold

ROGER setting

20 parents, 200 offspring, elitist selection (20 + 200)-ES

Crossover rate .6

Self adaptive mutation, rate 1.

Stopping criterion: 50,000 fitness evaluations

Bagged ROGER

20 runs on each fold

best hypothesis from each run retained

BH defined from these 20 hypotheses

Experimental results

1. The AUC criterion

Gender

AUC = .985

Activities

Activity	1105	1201	1401	2001	2002	3004	5102
AUC	0.96	0.99	0.86	0.80	0.89	0.73	0.94

From AUC optimization to classification

\mathbf{x}	x_1	x_2	x_3	\mathbf{e}	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
y	+1	+1	+1		-1	-1	+1	-1	+1	-1	-1	-1	-1
h	5.2	5.1	4.8	4.2	3.2	3	1.8	1.79	1.12	0.9	0.78	0.12	0.08

Define

$$p_h(y = 1|e) = \frac{\#\{\mathbf{x}_i, h(\mathbf{x}_i) < h(e), y_i = 1\}}{\#\{i, y_i = 1\}}$$

$$p_h(y = -1|e) = \frac{\#\{\mathbf{x}_i, h(\mathbf{x}_i) > h(e), y_i = -1\}}{\#\{i, y_i = -1\}}$$

Finally

$$h(e) = 1 \text{ iff } p_h(y = 1|e) > p_h(y = -1|e)$$

Here, $p_h(y = 1|e) = \frac{2}{5}$, $p_h(y = -1|e) = 0$

Experimental results

2. Misclassification rate

Gender

$Err(\text{Gender} = 1)$	$Err(\text{Gender} = 0)$
1.93%	0.90%

Activities

Activity A_i	1105	1201	1401	2001	2002	3004	5102
$Err(A = A_i)$	9.10%	2.18%	19.08%	30.72%	18.36%	32.56%	12.4%
$Err(A \neq A_i)$	8.11%	2.06%	18.32%	27.34%	16.85%	29.12%	12.77%

On unseen data

Vote of BH_1, \dots, BH_{10}

extracted from all 10 folds

First entry

All BH s have same weight

Second entry

$$\text{Weight}(BH_i) = \text{AUC}(BH_i)$$

More References on ROGER

- Impact Studies and Sensitivity Analysis in Medical Data Mining with ROC-based Genetic Learning.
M. Sebag and J. Azé and N. Lucas. 2003. Proceedings of IEEE International Conference on Data Mining, ICDM03, p. 637-640.
- Ensemble Feature Ranking,
K. Jong and J. Mary and A. Cornuejols and E. Marchiori and M. Sebag. 2004. In ECML/PKDD 2004, to appear.
- ROC-based Evolutionary Learning: Application to Medical Data Mining. M. Sebag and J. Azé and N. Lucas. 2004. Artificial Evolution'03, to appear. Springer Verlag LNCS.