Data grid models

Schedule

Alexis Bondu, Marc Boullé, Dominique Gay

January, 29, 2013

Schedule

- 14h15 : Data grid models
  - Principles, evaluation, optimisation
- 15h15 : Data Grid Models for Coclustering
  - Focus on model selection
- 16h15 : Pause (30 min)
- 16h45 : Coclustering applications using data grid models
  - Clustering of text, graph, text, curves, web logs…
- 17h15 : Data grid models for supervised learning
  - Application to data preparation and to change detection in stream mining
- 17h45 : Extension of data grid models
  - Classification rules and decision trees
- 18h30 : Conclusion
  - Summary, future work, discussion

Data grid models
Principles, evaluation, optimisation

Alexis Bondu, Marc Boullé, Dominique Gay

January, 29, 2013

Schedule

- Introduction
- Data grid models
- Applications
- Conclusion

Data table

instances x variables

Age	Education	Education Num	Marital status	Occupation	Race	Sex	Hours Per week	Native country	...	Class
39	Bachelors	13	Never-married	Adm-clerical	White	Male	40	United-States	...	less
50	Bachelors	13	Married-civ-spouse	Exec-managerial	White	Male	13	United-States	...	less
38	HS-grad	9	Divorced	Handlers-cleaners	White	Male	40	United-States	...	less
53	11th	7	Married-civ-spouse	Handlers-cleaners	Black	Male	40	United-States	...	less
28	Bachelors	13	Married-civ-spouse	Prof-specialty	Black	Female	40	Cuba	...	less
37	Masters	14	Married-civ-spouse	Exec-managerial	White	Female	40	United-States	...	less
49	9th	5	Married-spouse-absent	Other-service	Black	Female	16	Jamaica	...	less
52	HS-grad	9	Married-civ-spouse	Exec-managerial	White	Male	45	United-States	...	more
31	Masters	14	Never-married	Prof-specialty	White	Female	50	United-States	...	more
42	Bachelors	13	Married-civ-spouse	Exec-managerial	White	Male	40	United-States	...	more
37	Some-college	10	Married-civ-spouse	Exec-managerial	Black	Male	80	United-States	...	more
30	Bachelors	13	Married-civ-spouse	Prof-specialty	Asian	Male	40	India	...	more
23	Bachelors	13	Never-married	Adm-clerical	White	Female	30	United-States	...	less
32	Assoc-acdm	12	Never-married	Sales	Black	Male	50	United-States	...	less
...	...	...	...	...	...	...	...	...	...	...

Context

- Statistical learning
  - Objective: train a model
    - Classification: the output variable is categorical
    - Regression: the output variable is numerical
    - Clustering: no output variable
- Data preparation
  - Variable selection
  - Search for a data representation
- Importance of data preparation
  - For the quality of the results
  - 80% of the process time
  - Critical in case of large databases

Bottleneck

Objective
Towards an automation of data preparation

- Context
  - Statistical analysis of an instances*variables data table
- Objective
  - Variable subset selection method
  - Search for a data representation
- Evaluation criteria of the objective
  - Genericity
  - Parameter-free
  - Reliability
  - Accuracy
  - Interpretability
  - Efficiency

Proposed approach: MODL

- Data grid models for non parametric density estimation
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Data grid based on the cross-product of the univariate partitions, with a piecewise constant density estimation in each cell of the grid
  - Bayesian approach for model selection
  - Efficient optimization algorithms

Schedule

- Introduction
- Data grid models
- Applications
- Conclusion

Data grid models
for statistical analysis of a data table

- Output variables (Y) or input variables (X)
- Numerical or categorical variables
- From univariate to multivariate

	Univariate	Bivariate	Multivariate
Classification Y categorical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Regression Y numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \dots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \dots, Y_{K'} \mid X_1, X_2, \dots, X_K)$

Classification
Discretization of numerical variables

- Univariate analysis
  - Numerical input variable X
  - Categorical output variable Y

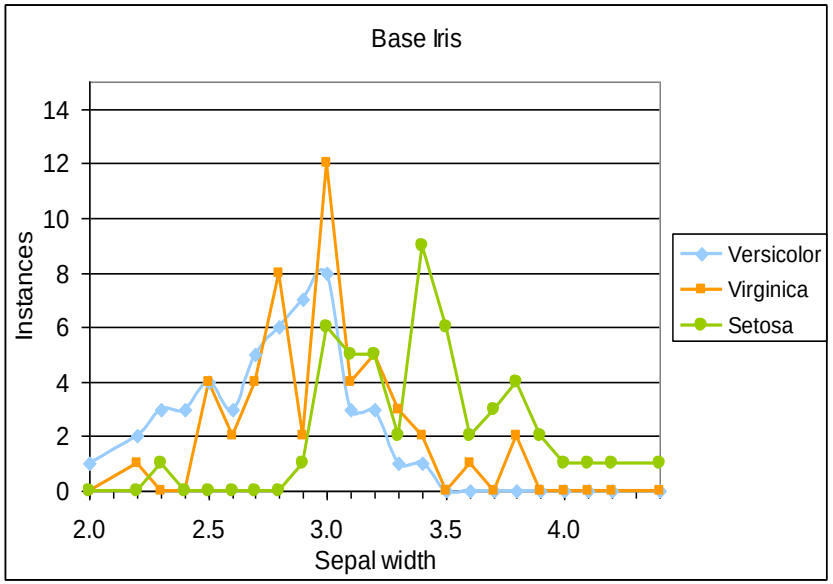
- Discretization

	Univariate	Bivariate	Multivariate
Classification **Y categorical**	**$P(Y \mid X)$**	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Regression Y numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \dots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \dots, Y_{K'} \mid X_1, X_2, \dots, X_K)$

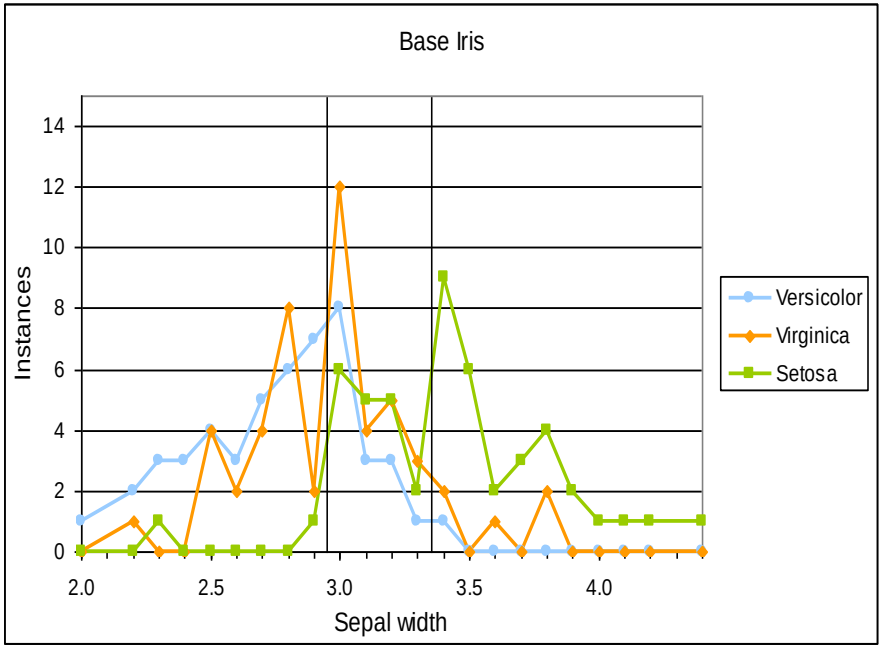
Numerical variables

Univariate analysis using supervised discretization

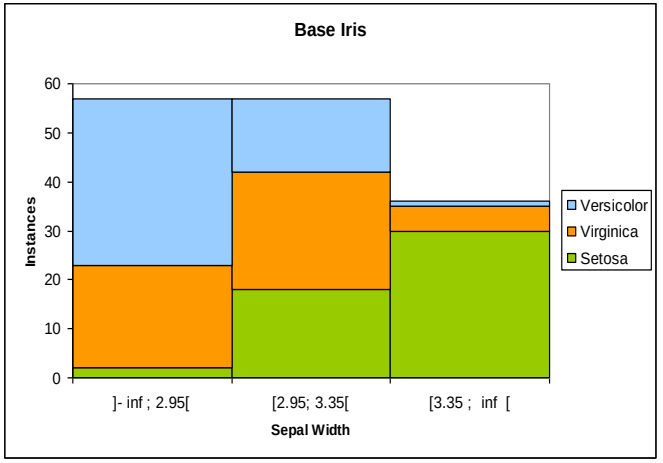
- Discretization:
  - Split of a numerical domain into a set of intervals
- Main issues:
  - Accuracy:
    - Good fit of the data
  - Robustness:
    - Good generalization



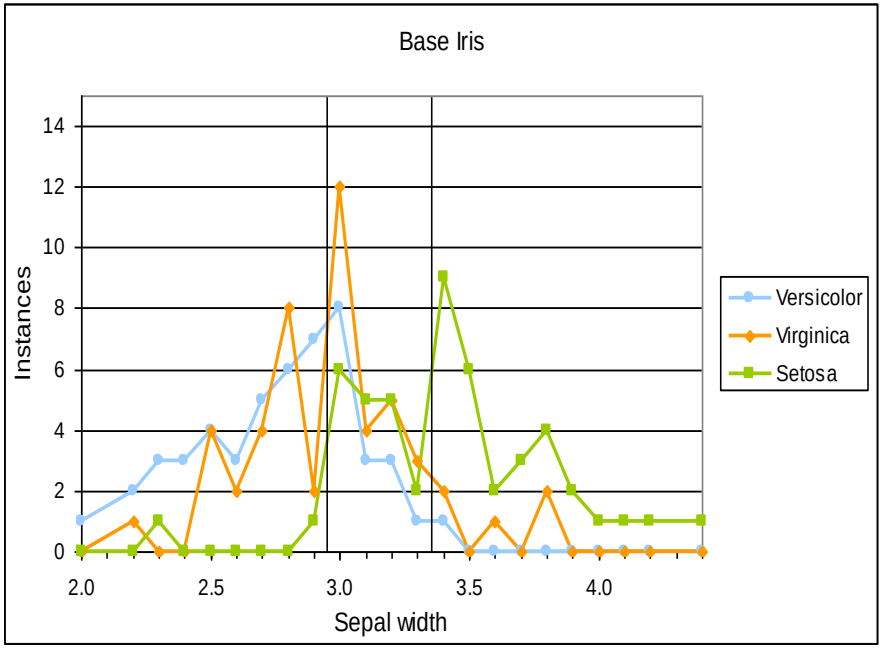
Supervised discretization
Model for conditional density estimation



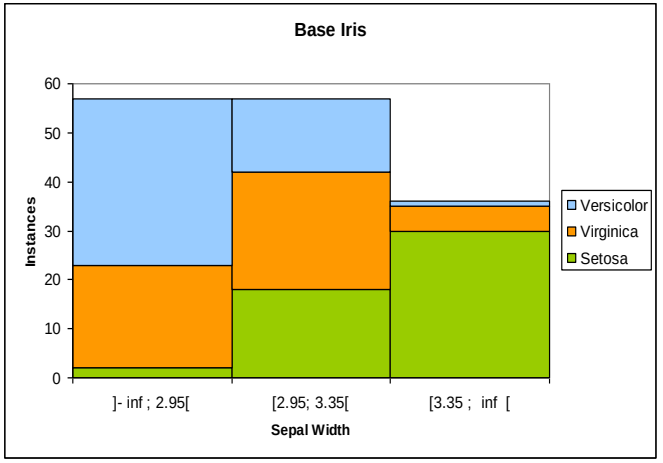
Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Sepal width	]- inf ; 2.95[	[2.95; 3.35[	[3.35 ; inf [



Supervised discretization
Model for conditional density estimation



Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Sepal width	]- inf ; 2.95[	[2.95; 3.35[	[3.35 ; inf [



How to select the best model?

Choice of rank statistics
Model of the sequence of the output values

- Example: sequence of 25 output values related to two classes ■ □

Choice of rank statistics
Model of the sequence of the output values

- Example: sequence of 25 output values related to two classes ■ □

- Discretization in one one interval
  - "Pure" model with one class

  - Mixture model

- Discretization in two intervals
  - Perfectly separable model

  - Partially separable model

How to select the best model?

Formalization

- **Definition:** A discretization model is defined by:
  - the number of input intervals,
  - the partition of the input variable into intervals,
  - the distribution of the output values in each interval.

Formalization

- **Definition:** A discretization model is defined by:
  - the number of input intervals,
  - the partition of the input variable into intervals,
  - the distribution of the output values in each interval.

- **Notations:**
  - N: number of instances
  - J: nombre of classes
  - I: number of intervals
  - $N_{i.}$: number of instances in the interval i
  - N_{ij}: number of instances in the interval i for class j

Bayesian approach for model selection

- Best model: the most probable model given the data
- Maximize $P(M \mid D) = \dfrac{P(M)P(D \mid M)}{P(D)}$
- Using a decomposition of the model parameters

$$P(M)P(D \mid M) = P(I)P(\{N_i\} \mid I)P(\{N_{ij}\} \mid I, \{N_i\})P(D \mid M)$$

- Assuming independence of the output distributions in each interval

$$P(M)P(D \mid M) = P(I)P(\{N_{i.}\} \mid I)\prod_{i=1}^{I} P(\{N_{ij}\} \mid I, \{N_{i.}\})\prod_{i=1}^{I} P(D_i \mid M)$$

- We now need to evaluate the prior distribution of the model parameters

Prior distribution of the models

- **Definition:** We define the hierarchical prior as follows:
  - the number of intervals is uniformly distributed between 1 et *N*,
  - for a given number of intervals *I*, every set of *I* interval bounds are equiprobable,
  - for a given interval, every distribution of the output values are equiprobable,
  - the distributions of the output values on each input interval are independent from each other.
- Hierarchical prior, uniformly distributed at each stage of the hierarchy

Optimal evaluation criterion MODL

- **Theorem**: A discretization model distributed according the hierarchical prior is Bayes optimal for a given set of instances if the following criterion is minimal:

$$\log(N) + \log\left(C_{N+I-1}^{I-1}\right) + \sum_{i=1}^{I} \log\left(C_{N_{i.}+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log\left(N_{i.}!/N_{i1}!N_{i2}!...N_{iJ}!\right)$$

prior | likelihood

- 1° term: choice of the number of intervals
- 2° term: choice of the bounds of the intervals
- 3° term: choice of the output distribution *Y* in each interval
- 4° term: likelihood of the data given the model

Discretization algorithm

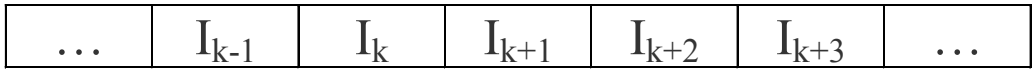
- Optimal solution in $O(N^3)$
  - Based on dynamic programing
  - Usefull to evaluate the quality of optimization heuristics

- Approximated solution in $O(N \log(N))$
  - Greedy bottom-up heuristic
    - 1) Initial solution: one interval per instance
    - 2) Evaluate all merges between adjacent intervals
    - 3) Perform best merge if improved criterion
    - 4) If improved criterion, repeat step 2, otherwise stop
  - Basic implémentation in $O(N^3)$
  - Efficient implémentation in $O(N \log(N))$
    - Exploiting the additivity of the criterion
    - Using a maintained sorted list of the best merges

Post-optimization of discretizations

Exhaustive search in a neighborhood of the current solution

Discretization intervals



Split of interval I_k

Merge of interval I_k and I_{k+1}

Merge-Split of intervals I_k and I_{k+1}

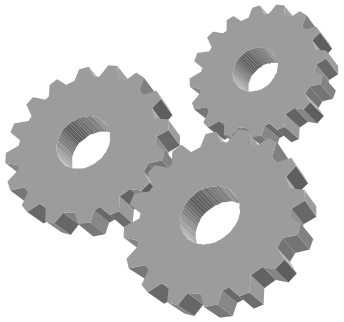
Merge-Merge-Split of intervals I_k, I_{k+1} and I_{k+2}

Quasi-optimal heuristic

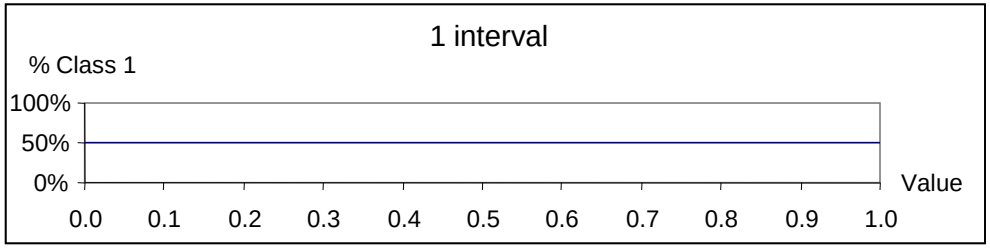
- Discretization algorithm MODL
  - Step 1: Greedy Merge
    - Iterative merges between intervals until no further improvement
  - Step 2: Exhaustive Merge
    - Iterative merges between intervals until one single global interval
    - Keep the best solution
  - Step 3: Post-optimization
    - Exhaustive search of local improvements in a neighborhood of the best solution
- Evaluation on 2000 discretizations
  - Optimal solution in more than 95% of the cases
  - In the remaining 5%, solution close from the optimal one (diff<0.15%)

Classification
Discretization of numerical variables

Evaluation

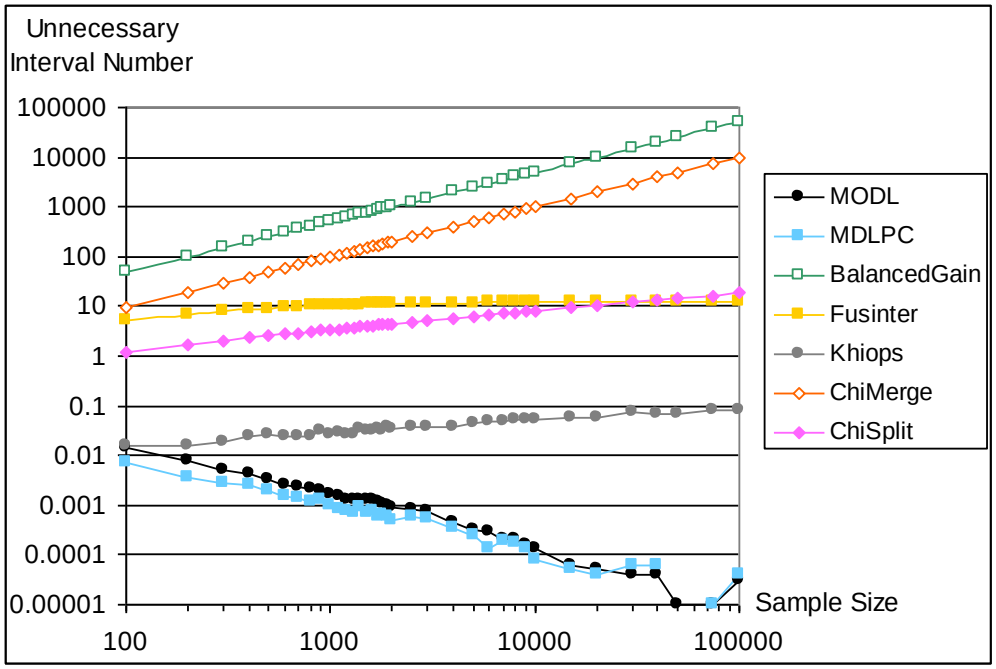


Discretization of a noise pattern

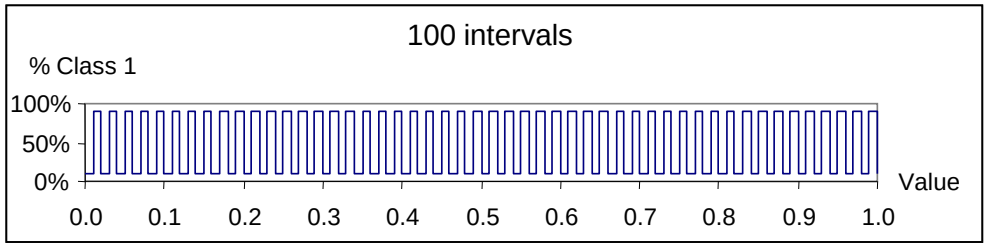


MODL reliably identifies the lack of predictive information

No over-fitting

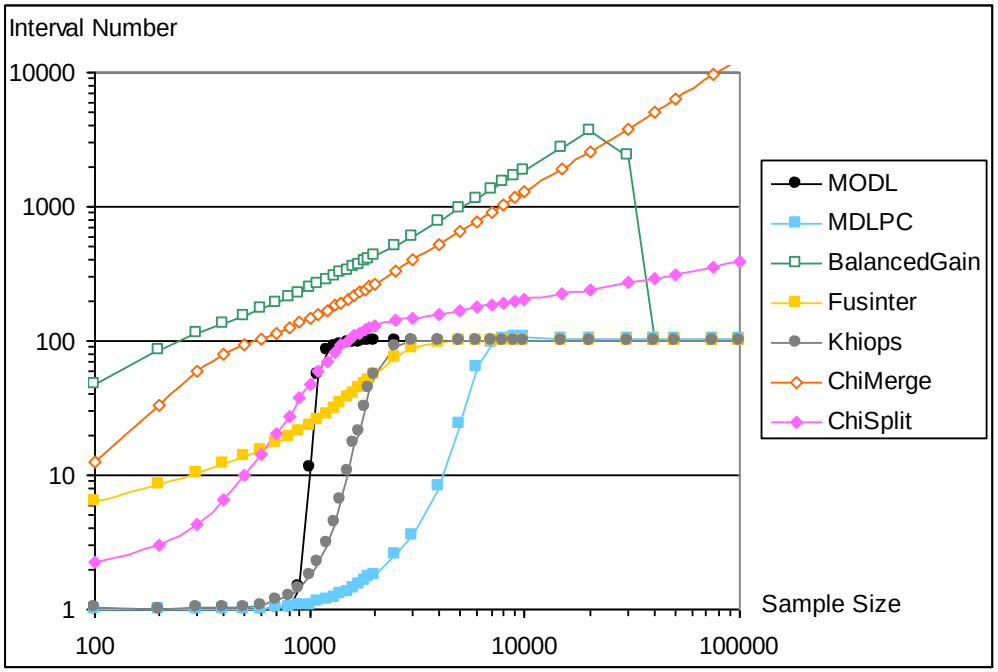


Discretization of a crenel pattern



MODL correctly identifies the relevant information with a minimal number of instances

No under-fitting



Classification
Value grouping of categorical variables

- Univariate analysis
  - Categorical input variable *X*
  - Categorical output variable *Y*

- Value grouping

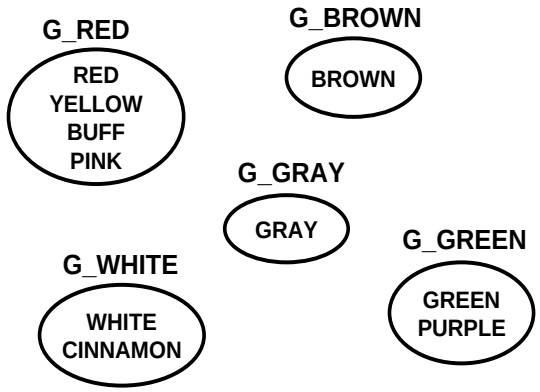
	Univariate	Bivariate	Multivariate
Classification ***Y* categorical**	**P(*Y* | *X*)**	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Regression *Y* numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \dots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \dots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \dots, Y_{K'} \mid X_1, X_2, \dots, X_K)$

Categorical variables

Univariate analysis using value grouping

Cap color	EDIBLE	POISONOUS	Frequency
BROWN	55.2%	44.8%	1610
GRAY	61.2%	38.8%	1458
RED	40.2%	59.8%	1066
YELLOW	38.4%	61.6%	743
WHITE	69.9%	30.1%	711
BUFF	30.3%	69.7%	122
PINK	39.6%	60.4%	101
CINNAMON	71.0%	29.0%	31
GREEN	100.0%	0.0%	13
PURPLE	100.0%	0.0%	10

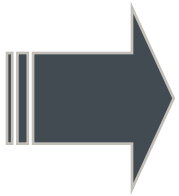
Cap color	EDIBLE	POISONOUS	Frequency
G_RED	38.9%	61.1%	2032
G_BROWN	55.2%	44.8%	1610
G_GRAY	61.2%	38.8%	1458
G_WHITE	69.9%	30.1%	742
G_GREEN	100.0%	0.0%	23



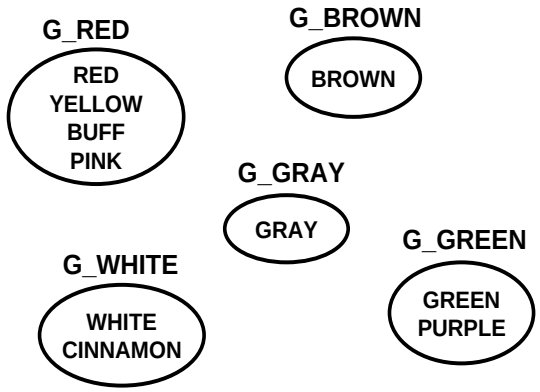
Categorical variables

Univariate analysis using value grouping

Cap color	EDIBLE	POISONOUS	Frequency
BROWN	55.2%	44.8%	1610
GRAY	61.2%	38.8%	1458
RED	40.2%	59.8%	1066
YELLOW	38.4%	61.6%	743
WHITE	69.9%	30.1%	711
BUFF	30.3%	69.7%	122
PINK	39.6%	60.4%	101
CINNAMON	71.0%	29.0%	31
GREEN	100.0%	0.0%	13
PURPLE	100.0%	0.0%	10



Cap color	EDIBLE	POISONOUS	Frequency
G_RED	38.9%	61.1%	2032
G_BROWN	55.2%	44.8%	1610
G_GRAY	61.2%	38.8%	1458
G_WHITE	69.9%	30.1%	742
G_GREEN	100.0%	0.0%	23



How to select the best model?

Value grouping

Same approach as for discretization

- A value grouping model is defined by:
  - the number of groups of inputs values,
  - the partition of the input variable into groups,
  - the distribution of the output values in each group.
- Model selection
  - Bayesian approach for model selection
  - Hierarchical prior for the model parameters
  - Exact analytical criterion to evaluate the models
- Optimization algorithms in $O(N \log(N))$

$$\log(V) + \log\left(B(V,I)\right) + \sum_{i=1}^{I} \log\left(C_{N_{i.}+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log\left(N_{i.}!/N_{i1}!N_{i2}!...N_{iJ}!\right)$$

prior | likelihood

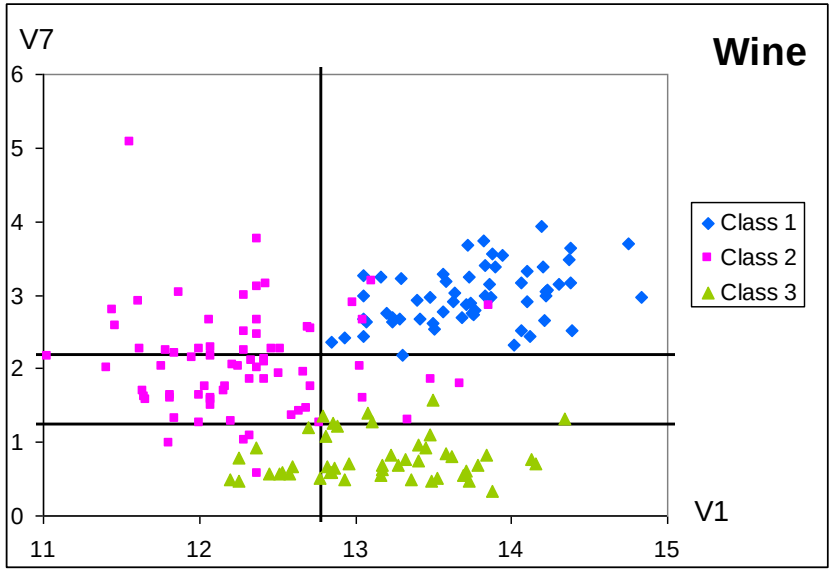
Classification supervisée
Bivariate discretization of numerical variables

- Bivariate analysis
  - Numerical input variables X_1 and X_2
  - Categorical output variable Y

- Bivariate discretization

	Univariate	**Bivariate**	Multivariate
Classification **Y categorical**	$P(Y \mid X)$	**$P(Y \mid X_1, X_2)$**	$P(Y \mid X_1, X_2, \ldots, X_K)$
Regression Y numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \ldots, Y_{K'} \mid X_1, X_2, \ldots, X_K)$

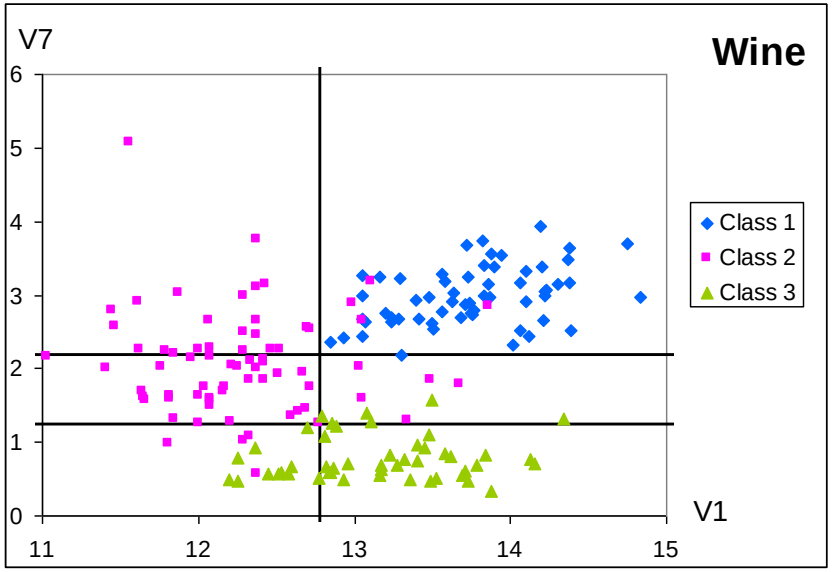
Pair of numerical variables



Pair of numerical variables

Bivariate discretization as a conditional density estimator

- Each input variable is discretized
- We obtain a bivariate data grid



- In each cell, the conditional density is estimated by counting

V7xV1	]-inf;12.78]	]12.78;+inf[
]2.18;+inf[	(0, 23, 0)	(59, 0, 4)
]1.235;2.18]	(0, 35, 0)	(0, 5, 6)
]-inf;1.235]	(0, 4, 11)	(0, 0, 31)

Application of the MODL approach

- Explicit formalization of the model family
  - Definition of the model parameters $I_1, I_2, N_{i_1..}, N_{.i_2.}, N_{i_1 i_2 j}$
- Definition of a prior distribution on the parameters of the bivariate discretization models
  - Hierarchical prior
  - Uniform distribution at each stage of the hierarchy
- We obtain an exact analytical evaluation criterion

prior

$$\log(N) + \log\left(C_{N+I_1-1}^{I_1-1}\right) + \log(N) + \log\left(C_{N+I_2-1}^{I_2-1}\right) + \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(C_{N_{i_1 i_2 .}+J-1}^{J-1}\right) +$$

likelihood

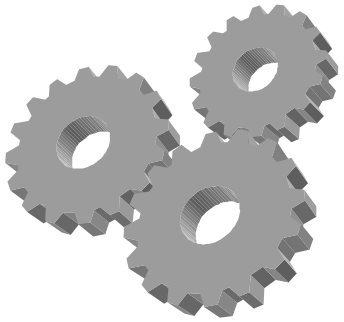
$$\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(N_{i_1 i_2 .}!/N_{i_1 i_2 1}!N_{i_1 i_2 2}!...N_{i_1 i_2 J}!\right)$$

Optimization algorithm

- Main algorithm: greedy bottom-up heuristic
- Post-optimization by alterning univariate optimizations
  - Freeze the partition of variable X_1 and optimize the partition of variable X_2
  - Freeze the partition of variable X_2 and optimize the partition of variable X_1
- Global optimization using the Variable Neighborhood Search (VNS) meta-heuristic
- Reduced algorithmic complexity: $O\left(N \log\left(N\right)\right)$
  - Exploit the sparseness of the data grid
    - At most N non-empty cells for N^2 potential cells
  - Exploit the additivity of the criterion

Classification
Bivariate discretization of numerical variables

Evaluation



Question: noise or information?

Diagram 1

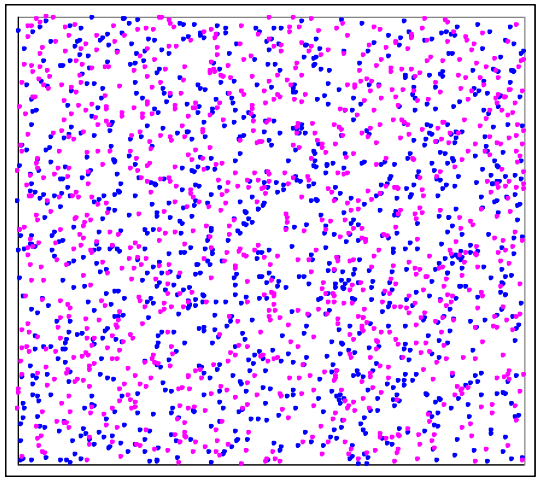


Diagram 2

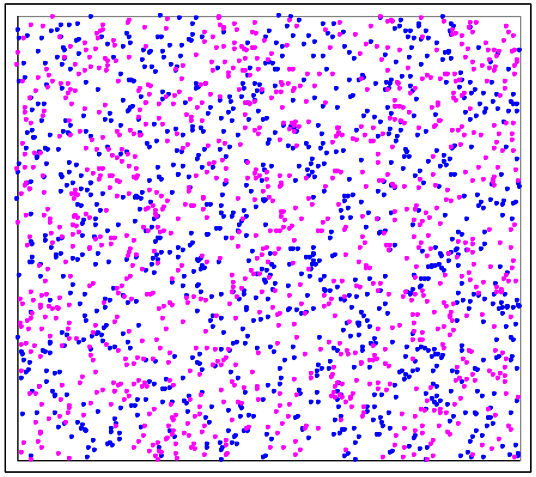


Diagram 3

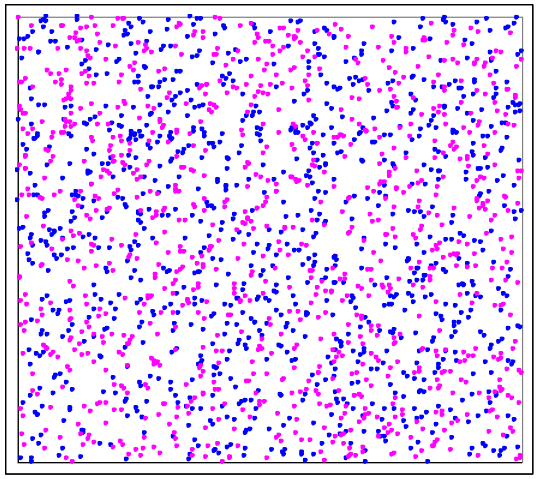
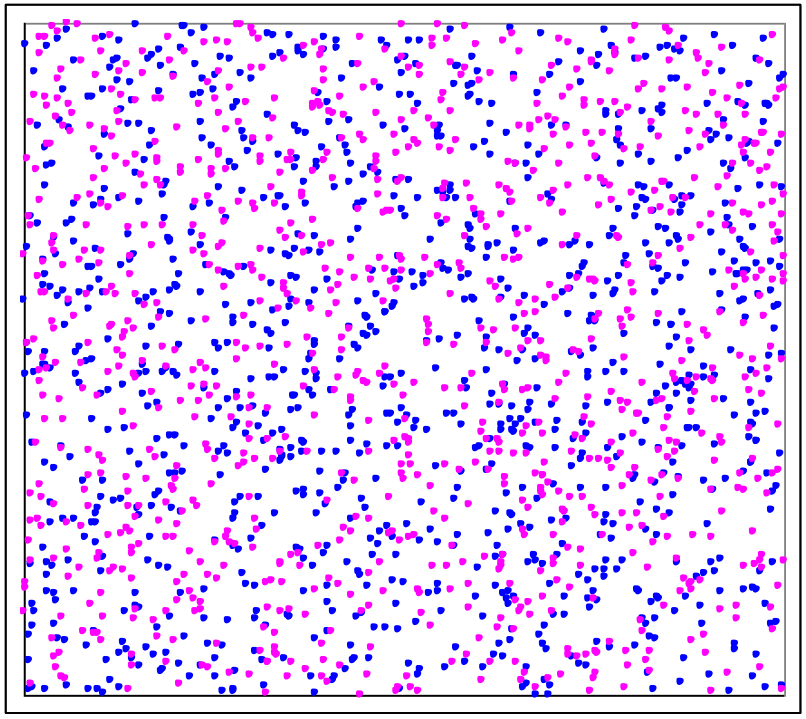
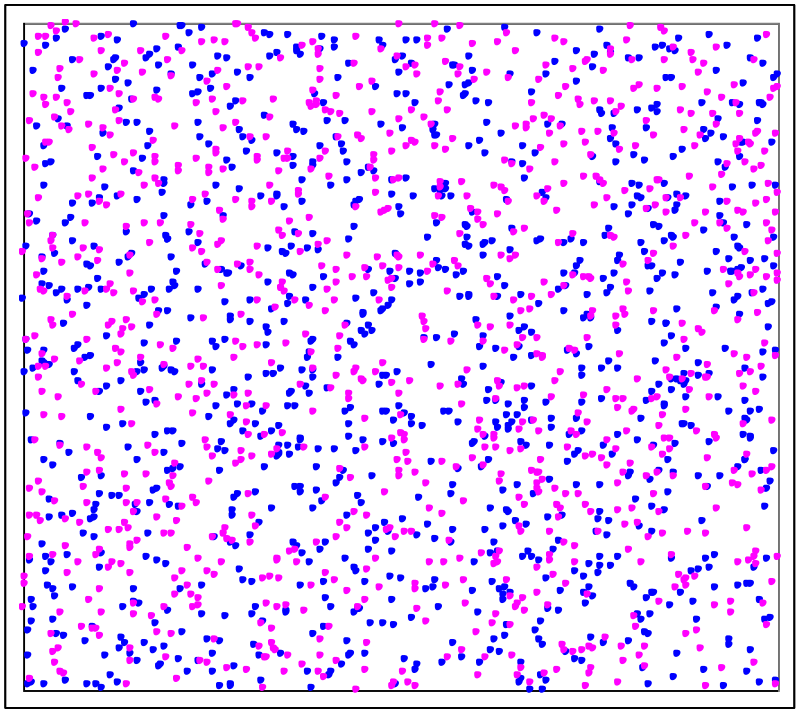


Diagram 1
noise

Diagram 1 (2000 points)



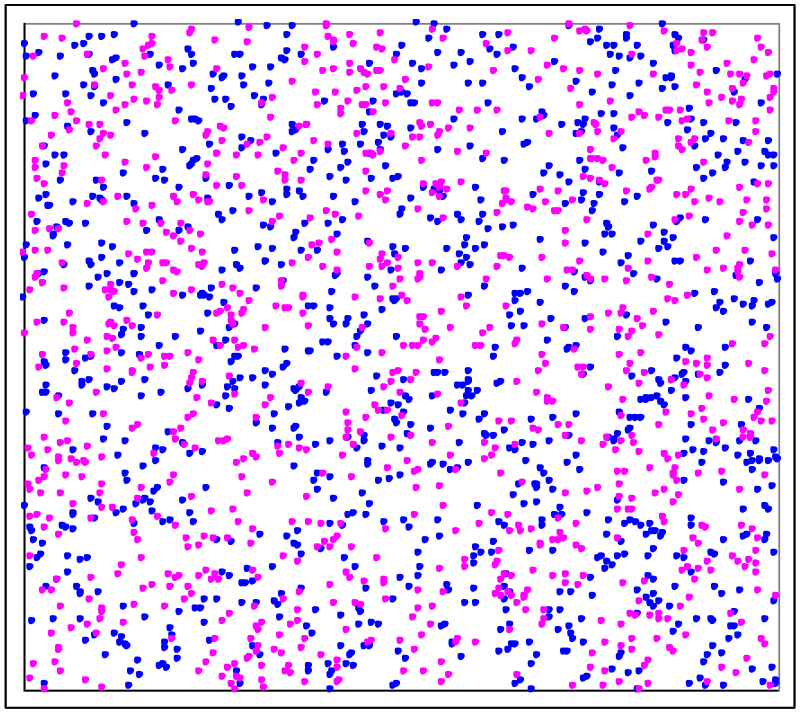
Data grid 1 x 1 (1 cell)



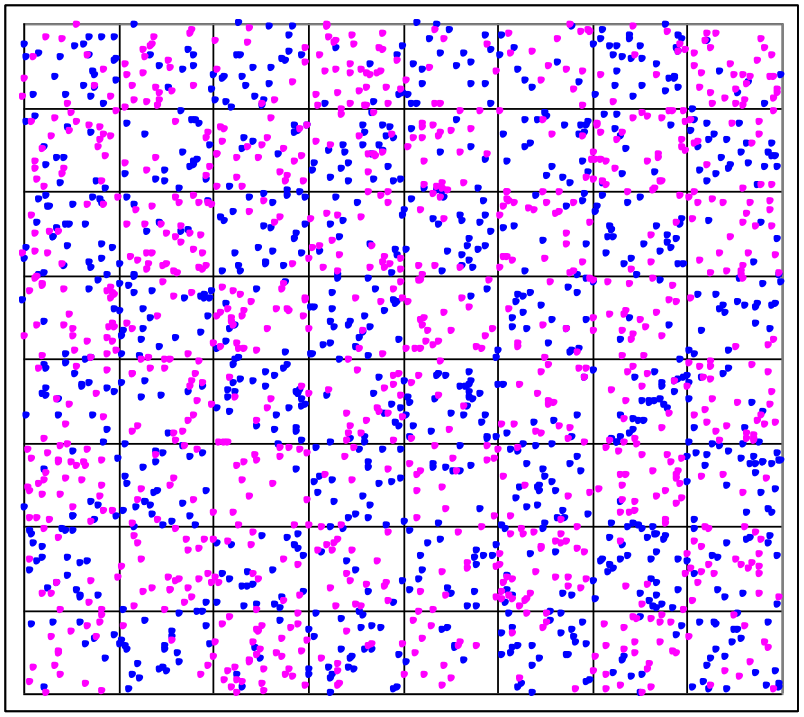
Value of criterion = 2075

Diagram 2
chessboard 8 x 8 with 25% noise

Diagram 2 (2000 points)



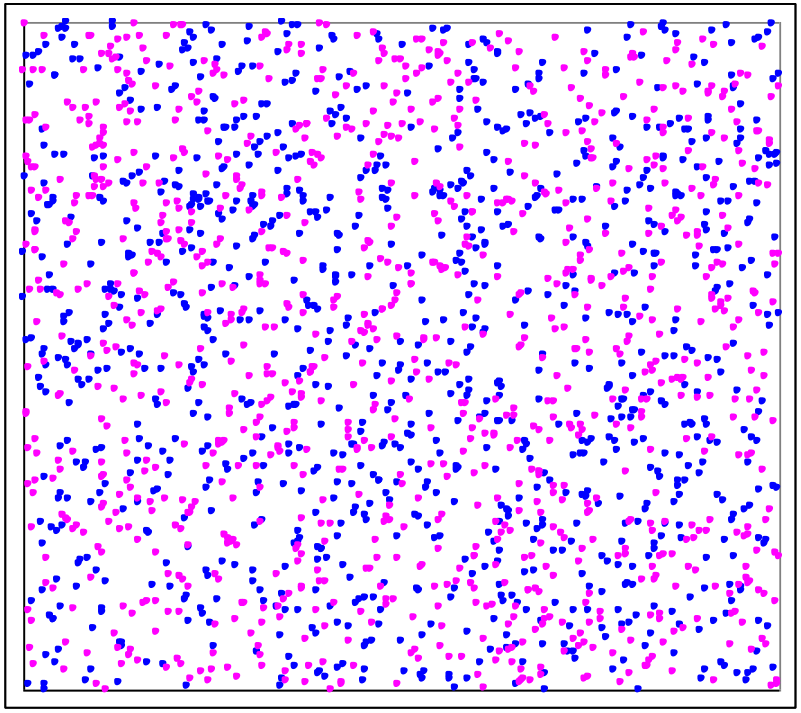
Data grid 8 x 8 (64 cells)



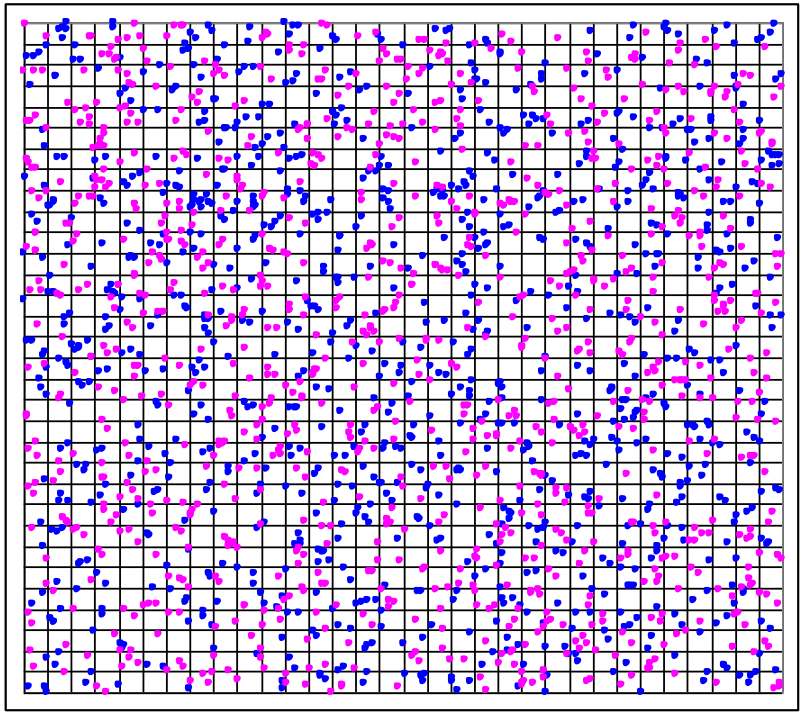
Value of criterion = 1900

Diagram 3
chessboard 32 x 32, without noise

Diagram 3 (2000 points)



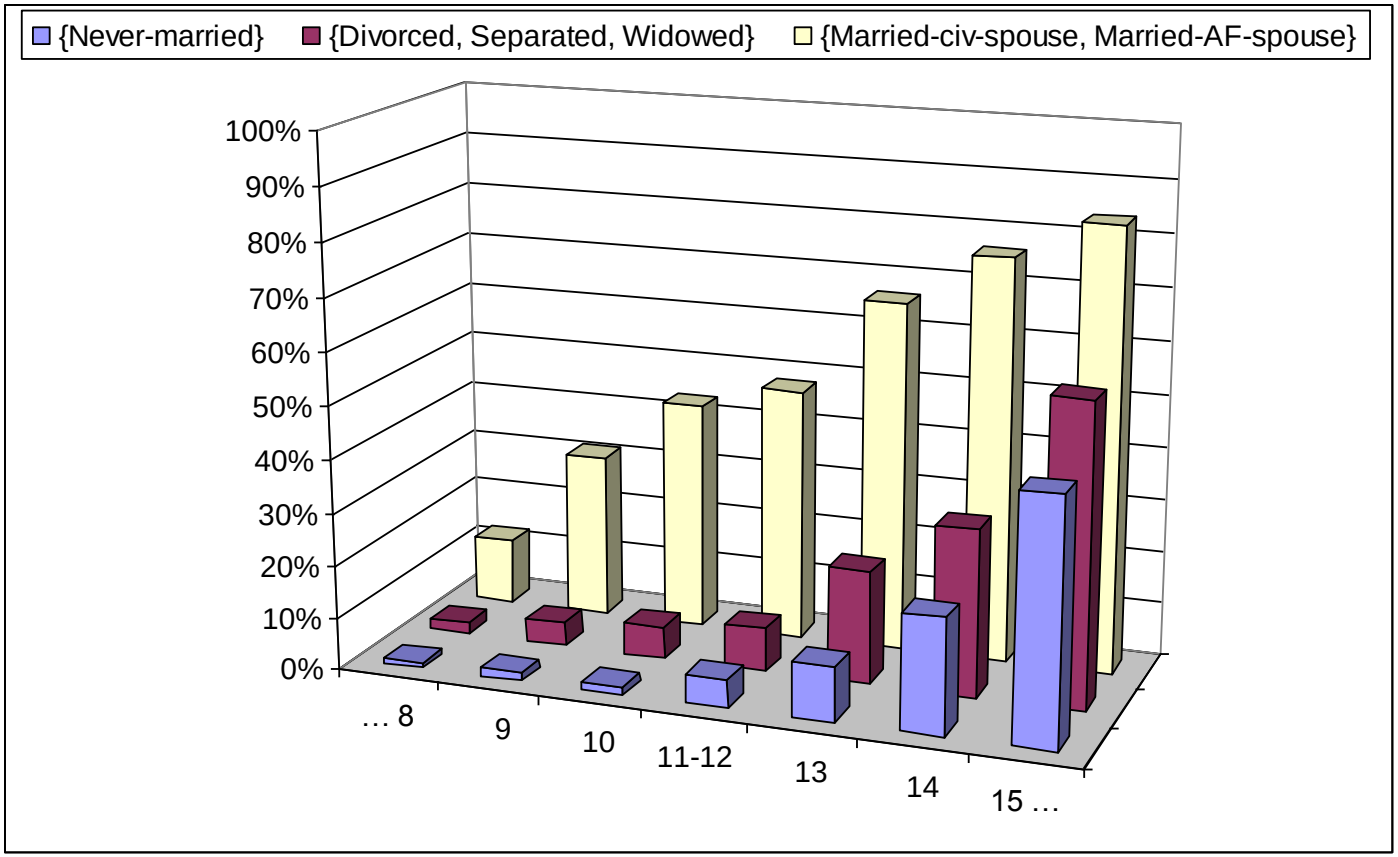
Data grid 32 x 32 (1024 cells)



Value of criterion = 1928

Visualization for a real database

Pair of numerical and categorical variables



Adult database

X_1: EducationNum

X_2: MaritalStatus

Y: Proportion of rich people

Classification
Multivariate data grid model

- Multivariate analysis
  - Numerical or categorical input variables X_1, X_2 … X_K
  - Categorical output variable Y

- Data grid model

	Univariate	Bivariate	**Multivariate**
Classification **Y categorical**	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	**$P(Y \mid X_1, X_2, \ldots, X_K)$**
Regression Y numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \ldots, Y_{K'} \mid X_1, X_2, \ldots, X_K)$

Multivariate data grid model for conditional density estimation

- Data grid model
  - Variable selection ($k \in \mathbf{K}_S$)
  - Discretization of numerical input variables ($k \in \mathbf{K}_1$)
  - Value grouping of categorical input variables ($k \in \mathbf{K}_2$)
  - For each cell of the related input data grid, description of the distribution of the output values
- Model selection
  - Bayesian approach for model selection
  - Hierarchical prior for model parameters
  - Exact analytical evaluation criterion

prior

$$\log(K+1) + \log\left(C_{K+K_S-1}^{K_S-1}\right) +$$

$$\sum_{k \in \mathrm{K}_S \cap \mathrm{K}_1} \left(\log(N) + \log\left(C_{N+I_k-1}^{I_k-1}\right)\right) +$$

$$\sum_{k \in \mathrm{K}_S \cap \mathrm{K}_2} \left(\log(V_k) + \log\left(B(V_k, I_k)\right)\right) +$$

$$\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_K=1}^{I_K} \log\left(C_{N_{i_1 i_2 \ldots i_K}+J-1}^{J-1}\right) +$$

likelihood

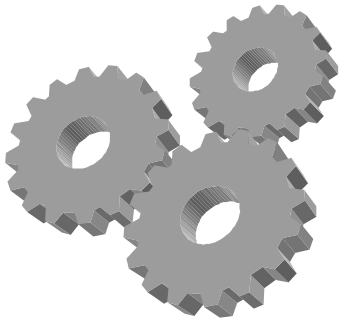
$$\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_K=1}^{I_K} \left(\log\left(N_{i_1 i_2 \ldots i_K}!\right) - \sum_{j=1}^{J} \log\left(N_{i_1 i_2 \ldots i_K j}!\right)\right)$$

Optimization of a multivariate data grid

- Complex problem
  - About $O((2^N)^K)$ possible data grid models (numerical case)
  - Each data grid contains $O(N^K)$ cells
  - Greed bottom-heuristic, with "naïve" implementation: $O(K^2 N^{K+2})$
- Optimization algorithm $O\left(KN\sqrt{N}\log(N)\max\left(K,\log(N)\right)\right)$
  - Exploit the sparseness of data grid models
    - At most N non-empty cells for N^K potential cells
  - Use incremental hashing functions
    - To efficiently retrieve non-empty cells
  - Exploit the additivity of the criterion

Classification
Multivariate data grid model

Evaluation



Adult database
Multivariate data grid as a classifier

- 6 selected variables, discretized or grouped
  - Data grid containing 252 non-empty cells
  - The 10 most frequent cells amount to 70% of the instances
  - Test accuracy: 86%

ID	Marital status	Education	Age	Hours per week	Capital gain	Capital loss	Effectif	% more
1	{Never-married, ...}	{Some-college, ...}	]28.5;+inf[	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	4957	2.3%
2	{Married,...}	{Some-college, ...}	]28.5;+inf[	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	4467	31.6%
3	{Never-married, ...}	{Some-college, ...}	]-inf;28.5]	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	4456	0.2%
4	{Married,...}	{Some-college, ...}	]28.5;+inf[	]40.5;+inf[	]-inf;1070.5]	]-inf;1551.5]	2511	41.4%
5	{Married,...}	{Doctorate, ...}	]28.5;+inf[	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	1698	60.2%
6	{Married,...}	{Doctorate, ...}	]28.5;+inf[	]40.5;+inf[	]-inf;1070.5]	]-inf;1551.5]	1518	72.7%
7	{Never-married, ...}	{Some-college, ...}	]28.5;+inf[	]40.5;+inf[	]-inf;1070.5]	]-inf;1551.5]	1422	8.0%
8	{Never-married, ...}	{Doctorate, ...}	]28.5;+inf[	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	1355	12.5%
9	{Never-married, ...}	{Preschool, ...}	]-inf;28.5]	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	1030	0.2%
10	{Married,...}	{Preschool, ...}	]28.5;+inf[	]-inf;40.5]	]-inf;1070.5]	]-inf;1551.5]	986	7.6%

Regression
Bivariate discretization of numerical variables

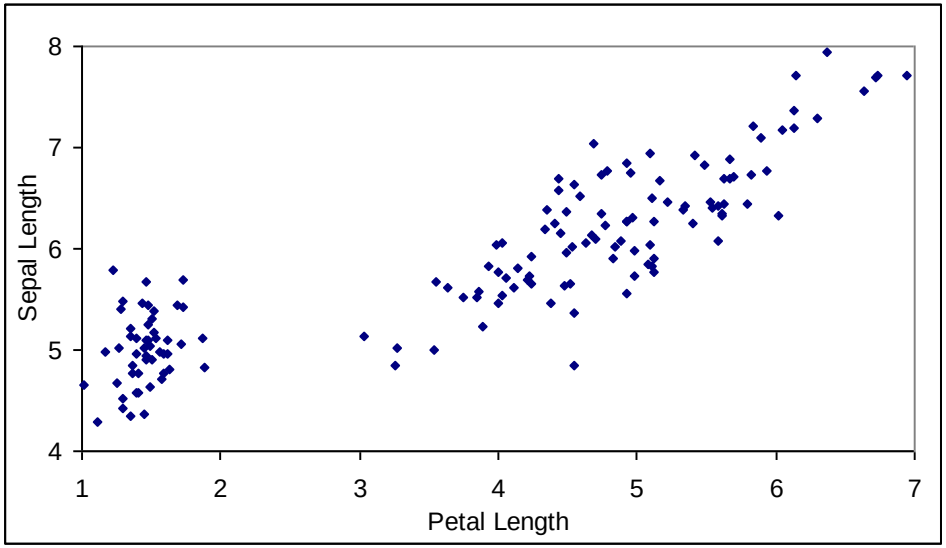
- Univariate analysis
  - Numerical input variable X
  - Numerical output variable Y

- Bivariate discretization of input and output variables

	Univariate	Bivariate	Multivariate
Classification Y categorical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Regression ***Y* numerical**	**$P(Y \mid X)$**	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \ldots, Y_{K'} \mid X_1, X_2, \ldots, X_K)$

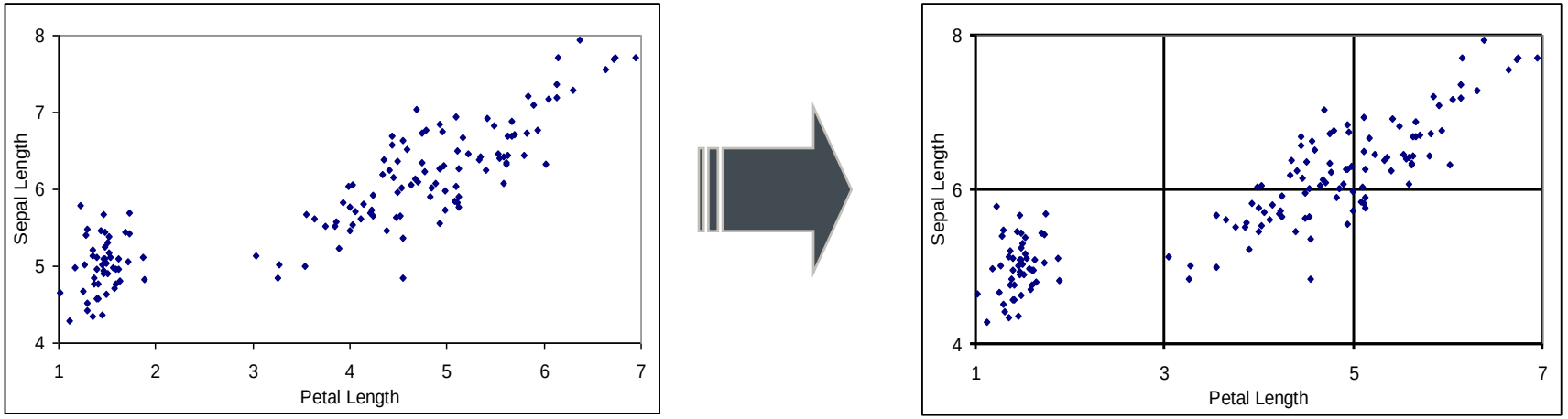
Numerical variables
Univariate analysis using a data grid model

- Prediction of the rank of *Y* given the rank of *X*
- Discretization:
  - Split a numerical domain into a set of intervals
- Main issues:
  - Accuracy:
    - Good fit of the data
  - Robustness:
    - Good generalization



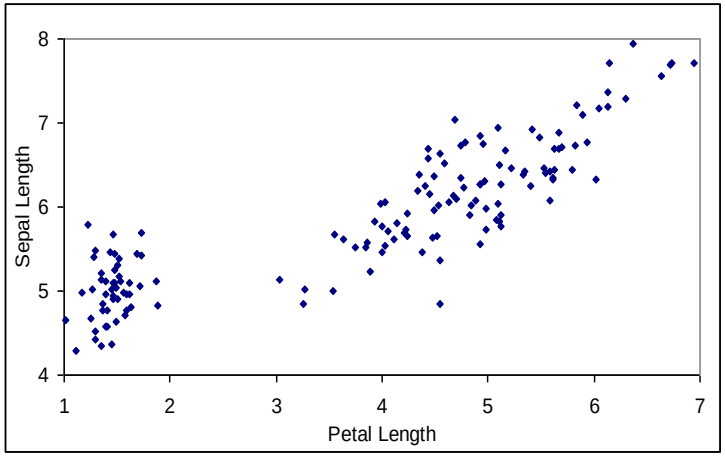
Bivariate discretization for regression

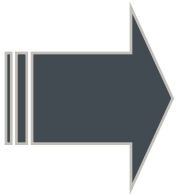
Data grid with 6 cells

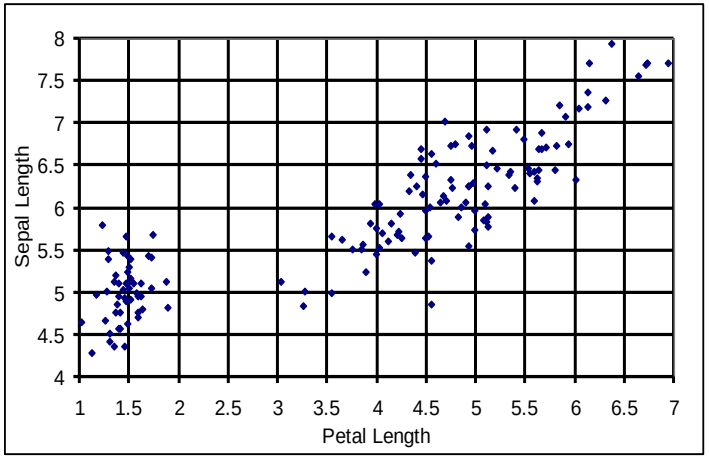


How to select the best model?

Bivariate discretization for regression
Data grid with 96 cells







How to select the best model?

Formalization

- **Definition:** A bivariate discretization model for regression is defined by: $M_N = \{I, J, \{N_{i.}\}, \{N_{ij}\}\}$
  - I: number of intervals for input variable X (between 1 and N)
  - J: number of intervals for output variable Y (between 1 and N)
  - $N_{i.}$: number of instances in input interval i
  - N_{ij}: number of instances in input interval i for output interval j
  - $N_{.j}$: number of instances in output interval j

- Bayesian approach for model selection

Exact analytical evaluation criterion

- Prior distribution of the model parameters :
  - Hierarchical prior
  - Independence between *I* and *J*
  - Independence between the output distribution in each input interval
  - Uniform prior at each stage of the hierarchy

- Two-stage likelihood:
  1. Probability of being in a given output interval,
  2. Probability of having a given rank locally to this output interval.

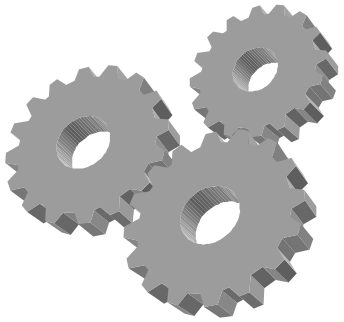
- Exact analytical criterion for the negative log of the posterior probability:

$$2\log(N)+\log\left(C_{N+I-1}^{I-1}\right)+\sum_{i=1}^{I}\log\left(C_{N_{i.}+J-1}^{J-1}\right)+\sum_{i=1}^{I}\log\left(N_{i.}!/N_{i1}!N_{i2}!...N_{iJ}!\right)+\sum_{j=1}^{J}\log\left(N_{.j}!\right)$$

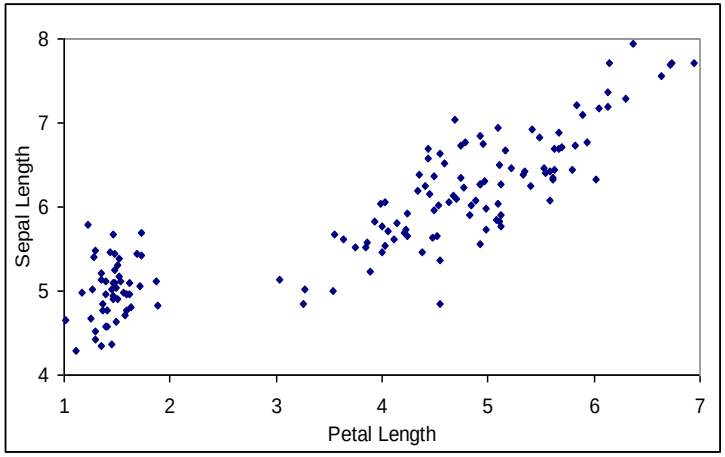
prior | likelihood

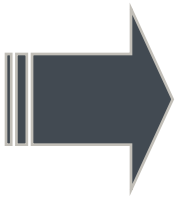
Regression
Data grid model

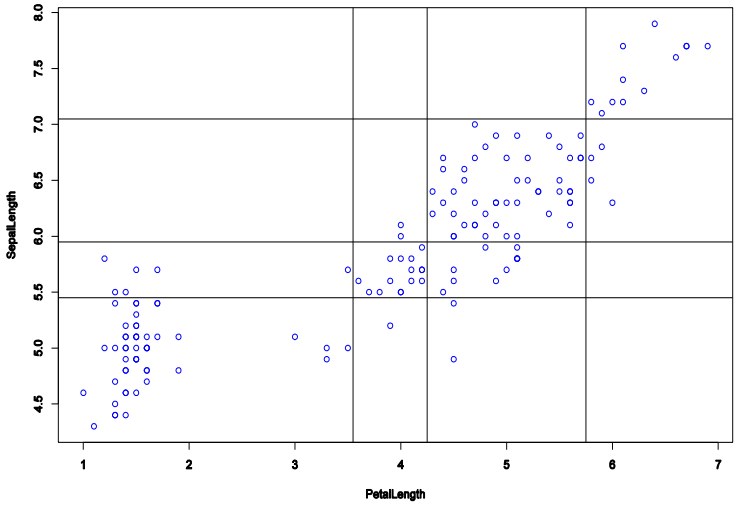
Evaluation



Bivariate discretization for regression
MODL optimal data grid

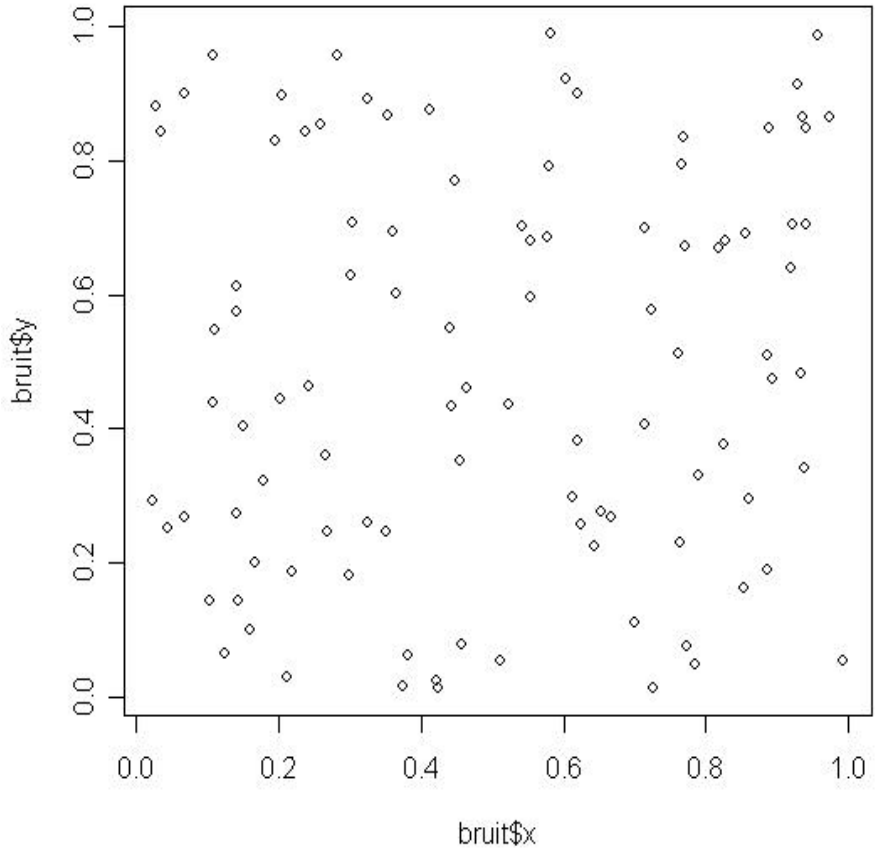






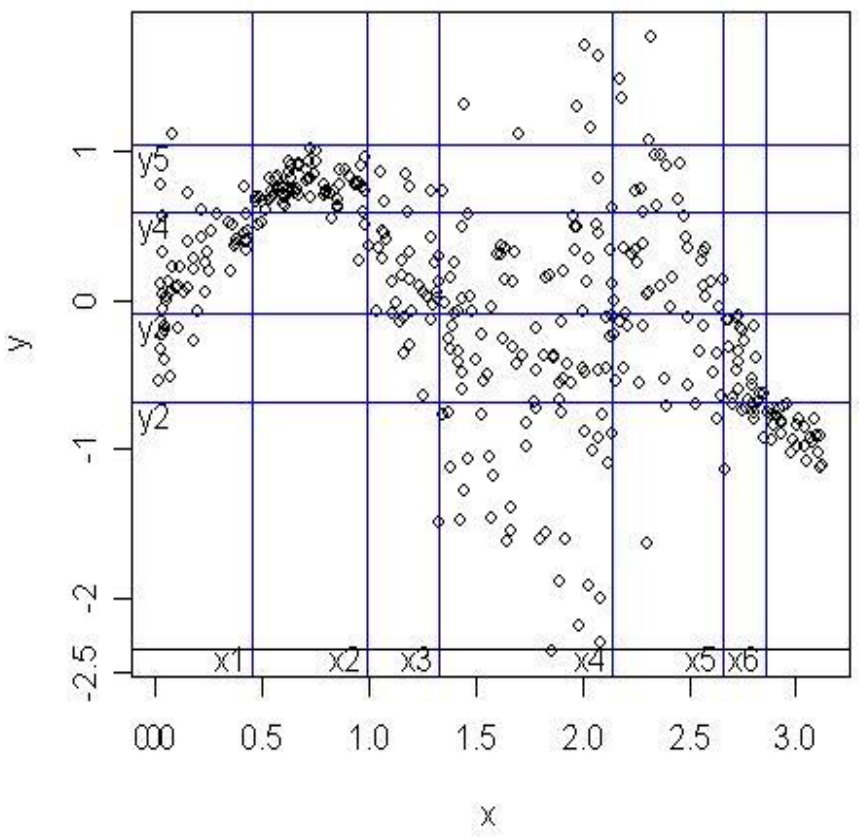
Bivariate discretization for regression

Dataset with no predictive information (noise)



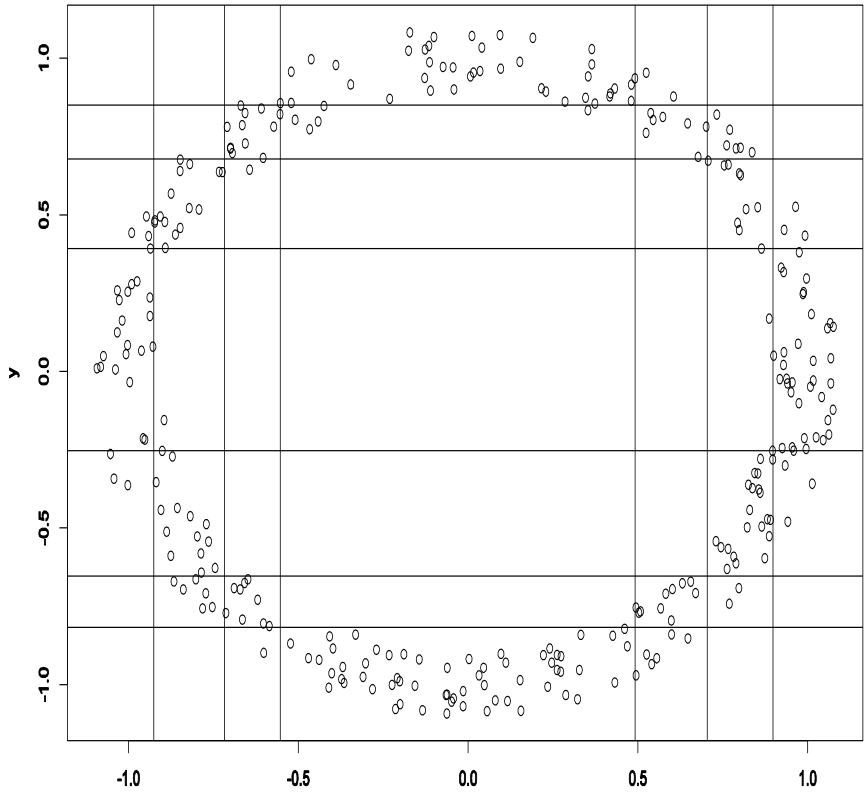
Bivariate discretization for regression

Dataset with heteroscedasticity (differing variance)



Bivariate discretization for regression

Dataset with multimodality



Data table instance x variables
Multivariate data grid model in the general case

- Multivariate analysis
  - Numerical or categorical input variables X_1, X_2 … X_K
  - Numerical or categorical output variables Y_1, Y_2 … $Y_{K'}$

- Data grid model

	Univariate	Bivariate	**Multivariate**
Classification Y categorical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Regression Y numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$
General case	–	–	$\mathbf{P(Y_1, Y_2, \ldots, Y_{K'} \mid X_1, X_2, \ldots, X_K)}$

Conditional density estimation
Data grid model

- Data grid model
  - Variable selection
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - We get an input data grid and an output data grid
  - For each cell of the input data grid, description of the distribution of the instances on the output cells

Variable selection

Data representation

- Model selection
  - Bayesian approach for model selection
  - Hierachical prior for the model parameters
  - Exact analytical criterion for the evaluation of the models

- Optimization algorithm $O\left(KN\sqrt{N}\log(N)\max\left(K,\log(N)\right)\right)$
  - Not yet implemented

Schedule

- Introduction
- Data grid models
- Applications
- Conclusion

Classification
Evaluation of univariate data grids

- Naive Bayes classifier
  - Univariate preprocessing MODL
  - Variable selection (MAP approach)
  - Compression-based model averaging
  - Chunking algorithms for large scale learning

- Challenges
  - Performance Prediction Challenge (IJCNN 2006)
    - 1st on two datasets (among five)
  - Causality Workbench Challenge (WCCI 2008)
    - 1st on four cases (among twelve)
  - Large Scale Learning Challenge (ICML 2008)
    - 1er on three datasets, 2nd on two datasets (among ten)

Classification
Evaluation of multivariate data grids

- Multivariate data grid model for classification
  - Model averaging
  - Coclustering instances-variables

- Challenges
  - Agnostic vs Prior Challenge (IJCNN 2007)
    - Agnostic track: 2nd et 3rd on two datasets (among five)
    - Prior track: 1st, 4th et 4th on three datasets (among five)

Regression
Evaluation of univariate data grids for regression

- Naive Bayes regressor
  - Univariate preprocessing MODL
  - Variable selection

- Challenges
  - Challenge "Evaluating Predictive Uncertainty" (IJCNN 2006)
    - 1st on two datasets (among five)

Use of Khiops in France Telecom

Khiops tools, available as a shareware http://www.khiops.com

- Marketing
  - Score team : scoring (churn, appetency, upselling, ...)
  - Fraud detection
  - PAC project: search of a representation among a very large number of variables
  - Analysis of survey data on the enterprise market
  - Analysis of the interest of the SIRET code on the entreprise market
  - Customer and product segmentation based on coclustering
  - Segmentation of customer verbatims
- Web mining
  - Orange Portail: detection of business queries
  - Web advertising
  - Detection of webspam
  - Segmentation of film reviews on a community site
- Text mining
  - Text classification
  - Segmentation of text corpus
  - Segmentation of SMS corpus
  - Recognition of named entities in text
- Network
  - Detection of peer-to-peer trafic based on analysis of IP packets
  - Network demand forecasting
  - Fading prediction in mobile transmission
- Purchasing service
  - Regression of landline and mobile telephone handsets
- Emotion mining
  - Detection of emotion from voice
  - Detection of emotion in SMS

Schedule

- Introduction
- Data grid models
- Applications
- Conclusion

MODL approach

- Density estimation using data grids
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Density estimation based on data grid models, with piecewise constant density per cell
  - Strong expressiveness
- Model selection
  - Bayesian approach for model selection
  - Hierarchical prior for the model parameters
  - Exact analytical criterion
- Optimization algorithm
  - Combinatorial algorithms
  - Heuristic exploiting the sparseness of the data grids and the additivity of the criterion
  - Efficient implementation

Genericity of the data grid models

	Univariate	Bivariate	Multivariate
Classification *Y* categorical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Regression *Y* numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$
General case	–	–	$P(Y_1, Y_2, \ldots, Y_{K'} \mid X_1, X_2, \ldots, X_K)$

MODL approach
Towards an automatisation of data preparation

- Advantages of the MODL approach
  - Genericity
  - Parameter-free
  - Reliability
  - Accuracy
  - Interpretability
  - Efficiency

- Khiops tool used in France Telecom in many contexts
  - Marketing, text mining, web mining, internet traffic, sociology, ergonomy…
  - Tool available as a shareware

http://www.khiops.com

Thank you for your attention!

EGC 2013 Tutorial – Data grid models

Data Grid Models for Coclustering

Focus on model selection

Alexis Bondu, Marc Boullé, Dominique Gay

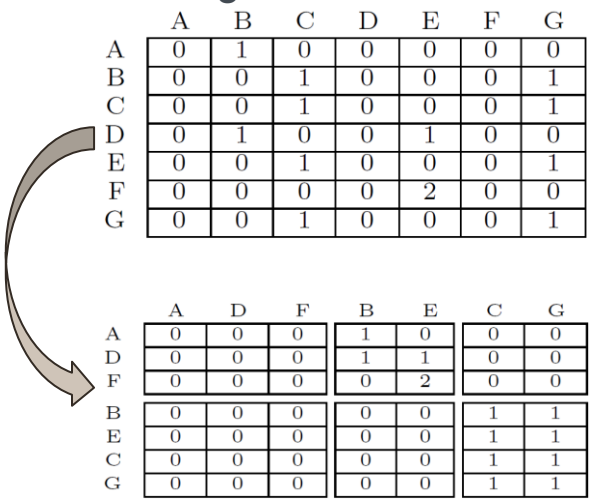
January, 29, 2013

Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

Coclustering

- Coclustering
  - Bi-clustering, bi-partitioning, co-partitioning, joint-partitioning, double clustering, coupled clustering, two-way clustering, block-modeling…
- Objective
  - Partition of the lines and rows of a matrix
  - Result in a partition of the matrix into coclusters
  - Maximize the homogeneity of the coclusters
- Type of matrix
  - Numerical matrix
  - Instances * variables matrix (numerical variables in most cases)
  - Contingency table



	A	B	C	D	E	F	G
A	0	1	0	0	0	0	0
B	0	0	1	0	0	0	1
C	0	0	1	0	0	0	1
D	0	1	0	0	1	0	0
E	0	0	1	0	0	0	1
F	0	0	0	0	2	0	0
G	0	0	1	0	0	0	1

	A	D	F	B	E	C	G
A	0	0	0	1	0	0	0
D	0	0	0	1	1	0	0
F	0	0	0	0	2	0	0
B	0	0	0	0	0	1	1
E	0	0	0	0	0	1	1
C	0	0	0	0	0	1	1
G	0	0	0	0	0	1	1

Coclustering of a contingency table
Bivariate Exploratory Analysis

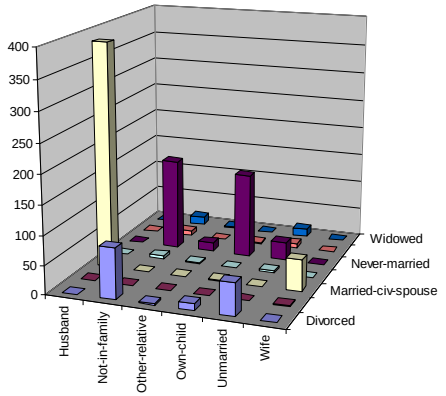
- Dataset
  - *N* instances
  - Two categorical variables *X* and *Y* having *V* and *W* values
  - Contingency table for bivariate analysis

relationship marital_status	Husband	Not-in-family	Other-relative	Own-child	Unmarried	Wife	Total
Divorced	0	86	3	12	55	0	156
Married-AF-spouse	0	0	0	0	0	1	1
Married-civ-spouse	385	0	0	2	0	54	441
Married-spouse-absent	0	5	1	1	3	0	10
Never-married	0	156	15	144	31	0	346
Separated	0	8	0	2	7	0	17
Widowed	0	14	2	0	13	0	29
Total	385	269	21	161	109	55	1000

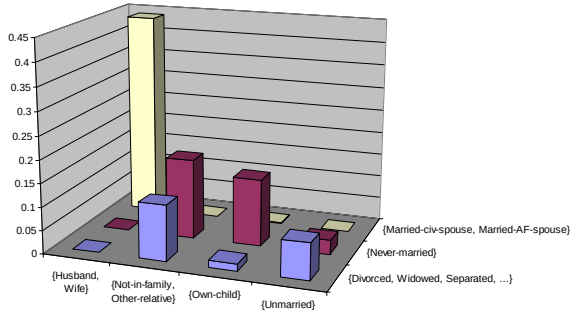
Coclustering

Contingency table after partitioning of the values

marital_status \ relationship	Husband	Wife	Not-in-family	Other-relative	Own-child	Unmarried	Total
Divorced	0	0	86	3	12	55	156
Married-spouse-absent	0	0	5	1	1	3	10
Separated	0	0	8	0	2	7	17
Widowed	0	0	14	2	0	13	29
Never-married	0	0	156	15	144	31	346
Married-AF-spouse	0	1	0	0	0	0	1
Married-civ-spouse	385	54	0	0	2	0	441
Total	385	55	269	21	161	109	1000



marital_status \ relationship	{Husband, Wife}	{Not-in-family, Other-relative}	{Own-child}	{Unmarried}	Total
{Divorced, Widowed, Separated, ...}	0	119	15	78	212
{Never-married}	0	171	144	31	346
{Married-civ-spouse, Married-AF-spouse}	440	0	2	0	442
Total	440	290	161	109	1000

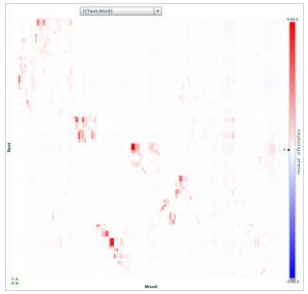


Applications of Coclustering

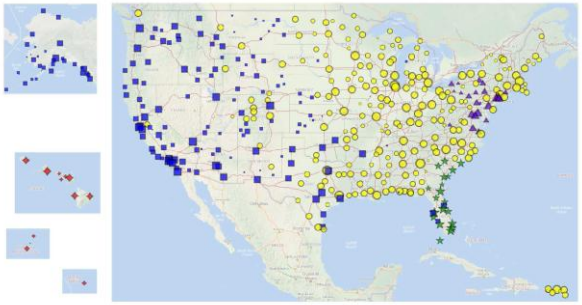
- Summary of the correlation between two categorical variables
  - Partition of the lines and rows of a contingency table
- Numerous applications
  - Text mining
    - Clusters of texts and clusters of words
  - Graph mining
    - Clusters of source and target vertices
  - Marketing
    - Cluster of customers and clusters of products
  - Web mining
    - Clusters of source and target web pages
  - Web usage mining
    - Clusters of cookies and clusters of web pages
  - …
- Potentially, large scale
  - Millions of instances
  - Tens of thousands of values

Applications of Coclustering
Some examples

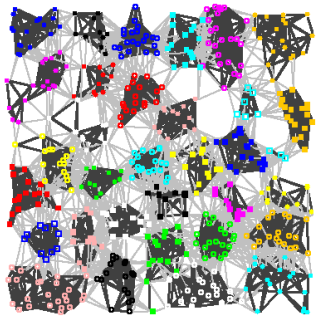
- 20 news-groups
  - Clusters of texts and word
  - Visualization of mutual information
  - Application to text categorization



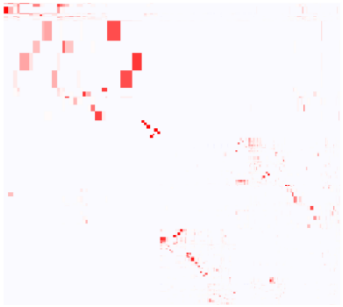
- US flight trips
  - Clusters of source and target airports
  - Visualization of the 5 main clusters of airports



- Graph mining
  - Clusters of source and target vertices



- Webspam
  - Clusters of source and target web pages
  - Visualization of mutual information
  - **Application to web spam detection**



Coclustering: main issues

- Assumptions regarding the distribution of the data
- Robustness
- Model parameters (cluster number…)
- Algorithm parameters
- Scalability

Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

Beyond clustering: coclustering

- Clustering
  - Group together similar individuals
  - Issues
    - Choice of a metric or similarity
    - Choice of the number of groups
    - Choice of an evaluation criterion and optimization algorithm
- Coclustering
  - First introduced by for the simultaneous clustering of the rows and columns of a numerical matrix
    - (Hartigan, 1972)
  - Applied for clustering of instances and variables
    - (Bock, 1979)

Quality of a coclustering
Some homogenity criteria

- Minimize the sum squared residue to approximate a numerical matrix
  - application to gene expression data (gene*condition)
    - (Cheng & Church, 2000, Cho et al., 2004)
- Minimize the loss of information for binary matrices
    - (Dhillon et al., 2003)
- Optimize an association measure between two variables
  - Euclidian distance between rows, or columns
    - method CREUC: (Govaert, 1983)
  - Khi2 criterion
    - method CROK12: (Govaert, 1983)
  - Goodman-Kruskal criterion
    - parameter-less method (Ienco et al., 2012)

Optimization
A combinatorial problem

- Number of partition of *n* elements into *k* non-empty subsets
  - Stirling numbers of the second kind $\left\{ {n \atop k} \right\}$

n \ k	0	1	2	3	4	5	6	7	8	9	10
0	1										
1	0	1									
2	0	1	1								
3	0	1	3	1							
4	0	1	7	6	1						
5	0	1	15	25	10	1					
6	0	1	31	90	65	15	1				
7	0	1	63	301	350	140	21	1			
8	0	1	127	966	1701	1050	266	28	1		
9	0	1	255	3025	7770	6951	2646	462	36	1	
10	0	1	511	9330	34105	42525	22827	5880	750	45	1

- Number of partition of *n* elements
  - Bell number B_n $\quad B_n = \sum_{k=0}^{n} \left\{ {n \atop k} \right\}$
    - $B_0, \ldots, B_9$: 1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147.
    - $B_{10} = 115\,975, \quad B_{20} = 51\,724\,158\,235\,372$
    - $B_{100} \sim 10^{116}, \quad B_{1000} \sim 10^{1927}, \quad B_{10000} \sim 10^{27644}, \ldots$

- Number of coclustering of a matrix *n*m*
  - $B_n * B_m$
  - Coclustering optimization: NP-complete problem!!!

Some optimization algorithms

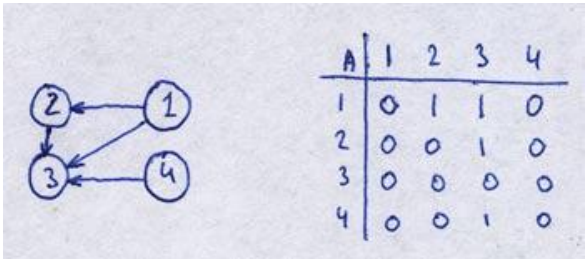
- Alternative optimization of each clustering
  - Freeze one clustering and optimize the other one
  - Reuse of standard clustering algorithms
    - K-means
    - Kohonen map
    - Bottom-up hierarchical clustering
    - Top-down divisive heuristic
    - …
- Local optimization of a coclustering of given size
    - Move values across clusters, for each dimension
- Global optimisation of a coclustering
  - Divisive approach: add one cluster per dimension at a time
  - Agglomerative approach: merge of clusters
  - Meta-heuristic to escape local optima
    - e.g.: simulated annealing

Limit of the standard approaches

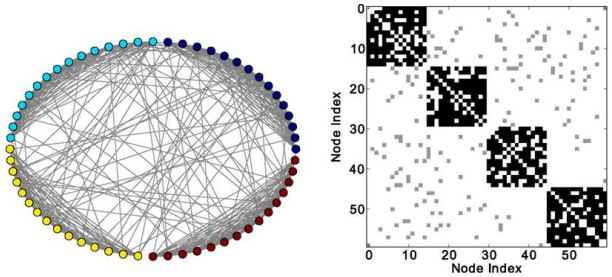
- Method requiring the numbers of clusters
  - Too small: under-fit the data
  - Too large: over-fit the data
- Parameter-less method
  - Optimization of a criterion such as Goodman-Kruskall
  - No regularization
  - Unknown robustness to noisy data
  - Problem of random data: no pattern
    - Need a threshold to avoid the detection of spurious patterns?

Block-modeling approaches
Relation to coclustering

- Approaches originating from social network analysis
  - A social network is a graph, represented by its adjacency matrix
    - number of edges from a source node to target node
  - Adjacency matrix: equivalent to a contingency table
    - Source node: first variable
    - Target node: second variable
    - Edge: individual



- Block-modeling
  - Cluster: group of actors having the same role
  - Block: type of relation between actors



Block-modeling approaches
Deterministic

- Early approaches: deterministic
  - (Lorrain & White, 1971), (Arabie et al., 1978)
  - Predefined type of block patterns
  - Use of adhoc functions measuring the fit of real blocks to predefined types of blocks
  - Standard clustering algorithms
  - Limits:
    - does not fit the stochastic nature of real world datasets

Block-modeling approaches
Stochastic

- Stochastic block-modeling
  - (Holland et al., 1983), (Wasserman et al., 1987), (Snijders & Nowicki, 1997), (Govaert & Nadif, 2003)
  - Bayesian approach with prior distribution on model parameters
    - Fixed number of clusters
    - Two latent variables for cluster assignments
    - Edge probabilities depend on the block
  - Non-parametric extension: Dirichlet process for partitions of any size
    - (Kemp et al., 2006)
  - Algorithm: Bayesian inference, alternate on partitions
    - EM, Gibbs sampling, MCMC, ML…
  - Limits:
    - assumptions regarding the distribution of the data
    - parameters (number of clusters, hyper-parameters for priors…)
    - scalability

MDL community detection
Minimum description length

- Minimum description length
  - (Rissanen, 1978), (Grünwald, 2007))
  - Two-part code: minimize code length of model + data given model
  - Parameter-less
  - Alleged resistance to over-fitting

- MDL coclustering methods
  - Early work: cross-association method directed simple graphs
    - (Chakrabarti et al., 2004)
    - Alternate optimization of clusters, plus divisive heuristic
  - Method for undirected simple graphs
    - (Rosvall & Bergstrom, 2007)
    - Optimization via simulated annealing
  - Study: importance of encoding scheme, coding node degrees
    - (Lang, 2009)
  - Limits:
    - Apply to directed or undirected simple graphs (binary contingency table)
    - Hidden prior assumptions behind the coding schema
    - Rely on empirical estimation of probabilities

Coclustering methods: synthesis

- Objective
  - Partition of the lines and rows of a matrix
- Type of matrix
  - Numerical matrix
  - Instances * variables matrix (numerical variables in most cases)
  - Contingency table
- Main approaches
  - Optimization of a quality measure
  - Statistical block-modeling
    - Bayesian model selection, MDL-based model selection
- Parameters
  - Number of clusters
  - Assumption regarding the distribution of the data (with potential prior parameters)
  - Parameter-less in some cases
- Optimized criterion
  - Quality measure
  - Posterior probability
  - Code length (MDL)
- Optimization algorithmic
  - Greedy bottom-up (divisive) or top-down (agglomerative) heuristic
  - Local optimization of clusters by moving values across clusters
  - Meta-heuristic: simulated annealing…
  - Bayesian optimization: Gibbs sampling, MCMC…

Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

Unsupervised data grids

- Data grid models for non parametric joint density estimation
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Data grid based on the cross-product of the univariate partitions, with a piecewise constant density estimation in each cell of the grid
  - Bayesian approach for model selection
  - Efficient optimization algorithms

- Coclustering
  - Focus on the case of two categorical variables
  - Extension available for multiple numerical or categorical variables

Coclustering of a contingency table
Bivariate Exploratory Analysis

- Dataset
  - *N* instances, two categorical variables *X* and *Y* having *V* and *W* values

How to chose the best model?

marital_status \ relationship	Husband	Wife	Not-in-family	Other-relative	Own-child	Unmarried	Total
Divorced	0	0	86	3	12	55	156
Married-spouse-absent	0	0	5	1	1	3	10
Separated	0	0	8	0	2	7	17
Widowed	0	0	14	2	0	13	29
Never-married	0	0	156	15	144	31	346
Married-AF-spouse	0	1	0	0	0	0	1
Married-civ-spouse	385	54	0	0	2	0	441
Total	385	55	269	21	161	109	1000

marital_status \ relationship	{Husband, Wife}	{Not-in-family, Other-relative}	{Own-child}	{Unmarried}	Total
{Divorced, Widowed, Separated, ...}	0	119	15	78	212
{Never-married}	0	171	144	31	346
{Married-civ-spouse, Married-AF-spouse}	440	0	2	0	442
Total	440	290	161	109	1000

Coclustering using data grid

- Principles
  - Define a model space
    - Model of the finite data sample
  - Define a prior on this model space
    - Exploits the hierarchy of the parameters
    - Uniform at each stage of the hierarchy
    - Data dependent prior
  - Obtain an exact analytical evaluation criterion (MAP)
  - Exploit scalable optimization heuristics

- Main features
  - Non-parametric
    - No assumptions regarding the distribution of the data
    - Universal estimator of joint density
    - Asymptotical convergence to the true joint density
  - Robustness
    - Do not over-fit the data
  - Parameter-free
  - Scalable

MODL Approach
Model definition

- **Definition:** A coclustering model is defined by:
  - a number of groups for each variable,
  - the partition of each variable into groups of values,
  - the multinomial distribution of the instances on the cells of the data grid resulting from the Cartesian product of the partitions of the variables,
  - the multinomial distribution of the instances of each group on the values of the group, for each variable.

- Notation:
  - X, Y — categorical variables
  - N — number of instances of the data sample
  - V, W — number of values for each variable
  - I, J — number of groups for each variable
  - $G=IJ$ — number of cells in the data grid
  - $i(v)$, $j(w)$ — index of group for value v (resp. w)
  - $m_{i.}$, $m_{.j}$ — number of values of group i (resp. j)
  - $n_{v.}$, $n_{.w}$ — number of instances for value v (resp. w)
  - n_{vw} — number of instances for pair of values (v, w)
  - $N_{i.}$, $N_{.j}$ — number of instances in group i (resp. j)
  - N_{ij} — number of instances in cell (i, j) of the data grid

MODL Approach
Bayesian approach for model selection

- Best model: the most probable model given the data
  - Maximum a Posteriori (MAP)

- Maximize $$P(M \mid D) = \frac{P(M)P(D \mid M)}{P(D)}$$

- Minimize $$-\log P(M \mid D) = -\log P(M) - \log P(D \mid M)$$
  - Bayesian interpretation
    - $P(M)$: prior
    - $P(D|M)$: likelihood
  - MDL interpretation
    - -log $P(M)$: coding length of model parameters
    - -log $P(D|M)$: coding length of data given model

MODL Approach
Prior distribution of model parameters

- Definition: The hierarchical prior on coclustering models is defined as follows:
  - the numbers of groups of values *I* (resp. *J*) are independent from each other, and uniformly distributed between 1 and *V* (resp. *W*),
  - for a given number of groups *I* of *X* (resp. *J* of *Y*), every partition of the *V* (resp. *W*) values into *I* (resp. *J*) groups is equiprobable,
  - for a data grid of given size (*I*, *J*), every multinomial distribution of the *N* instances on the *G=I*J* cells of the data grid is equiprobable,
  - for a given group of values of a given variable, every multinomial distribution of the instances of the group on the values of the group is equiprobable.

- Hierarchical prior, uniform at each level of the hierarchy

MODL Approach
Optimal evaluation criterion

- **Theorem**: The negative log of the posterior probability of a coclustering model *M* distributed according to the hierarchical prior is given by the following criterion

$$c(M) = \log V + \log W + \log B(V, I) + \log B(W, J)$$
$$+ \log \binom{N+G-1}{G-1}$$
$$+ \sum_{i=1}^{I} \log \binom{N_{i.}+m_{i.}-1}{m_{i.}-1} + \sum_{j=1}^{J} \log \binom{N_{.j}+m_{.j}-1}{m_{.j}-1}$$
$$+ \log N! - \sum_{i=1}^{I}\sum_{j=1}^{J} \log N_{ij}!$$
$$+ \sum_{i=1}^{I} \log N_{i.}! + \sum_{j=1}^{J} \log N_{.j}! - \sum_{v=1}^{V} \log n_{v.}! - \sum_{w=1}^{W} \log n_{.w}!$$

Detailed next slide

MODL analytical criterion

Exact evaluation of the posterior probability of a model

Prior

Coding length of model parameters

Likelihood

Coding length of data given model

$$c(M) = \log V + \log W$$

Number of groups for each variable

$$+ \log B(V, I) + \log B(W, J)$$

Partition of values into groups for each variable

$$+ \log \binom{N + G - 1}{G - 1}$$

Multinomial distribution of the instances of the dataset on the cells of the data grid

$$+ \sum_{i=1}^{I} \log \binom{N_{i.} + m_{i.} - 1}{m_{i.} - 1}$$

Multinomial distribution of the instances of each group on the values of the group (X_1)

$$+ \sum_{j=1}^{J} \log \binom{N_{.j} + m_{.j} - 1}{m_{.j} - 1}$$

Multinomial distribution of the instances of each group on the values of the group (X_2)

$$+ \log N! - \sum_{i=1}^{I} \sum_{j=1}^{J} \log N_{ij}!$$

Position of the instances among the cells according to the multinomial distribution

$$+ \sum_{i=1}^{I} \log N_{i.}! - \sum_{v=1}^{V} \log n_{v.}!$$

Value of the instances among the group values according to the multinomial distribution (X_1)

$$+ \sum_{j=1}^{J} \log N_{.j}! - \sum_{w=1}^{W} \log n_{.w}!$$

Value of the instances among the group values according to the multinomial distribution (X_2)

Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

Another interpretation
Non parametric estimation of joint density

- A coclustering model is a non parametric joint density estimator
  - Joint density $p(v,w) = p(X = v, Y = w)$
  - Joint density estimation according to a coclustering model with granularity (I, J)

$$p_{vw} = \frac{N_{ij}}{N} \frac{n_{v.}}{N_{i.}} \frac{n_{.w}}{N_{.j}}$$

- Two extreme cases
  - Null model $M_{\emptyset}$
    - Coarsest granularity (I, J) =(1,1)
    - Independence between variables

$$p_{vw} = \frac{n_{v.}}{N} \frac{n_{.w}}{N}$$

  - Maximal model M_{Max}
    - Finest granularity (I, J) =(V,W)
    - Empirical evaluation of joint probability
    - All correlation captured

$$p_{vw} = \frac{n_{vw}}{N}$$

Modeling with data dependent prior

- The model space is data dependent
  - The model space is discrete: we consider distributions of exactly *N* instances on a data grid of each granularity (*I*, *J*)
  - The number of model parameters increases with the sample size
  - Parameters have discrete values, number of parameter values is data dependent
- The prior is data dependent
  - The number of considered multinomial distributions depends on the size of the data sample

$$C_{N+IJ-1}^{IJ-1}$$

- The likelihood is discrete
  - The likelihood is null if the observed distribution is not exactly that of the model
  - Otherwise, it is based on the number of ways of observing the data with the exact multinomial counts

$$\frac{N!}{\prod_{i=1}^{I}\prod_{j=1}^{J} N_{ij}!}$$

- What is modeled ?
  - Direct modeling of the data sample using discrete distributions, instead of real-valued distributions

Modeling with data dependent prior
Main features

- Natural prior in finite size model space
  - Hierarchical, uniform at each level
  - ≠ Bayesianism: avoid uneasy choice of prior and hyper-parameters
    - parametric: e.g. Dirichlet distribution
    - non parametric: e.g. Dirichlet or Pitman-Yor process
  - ≠ MDL:
    - avoid uneasy choice of coding schemes
    - avoid empirical entropy estimation with asymptotical validity
  - Reduces to counting
- Easy estimation of likelihood
  - Reduces to counting
- Analytical formula for posterior probability of models
- Combinatorial algorithms
  - Main heuristic: greedy bottom-up merge heuristic, with guaranteed time complexity $O\left(N\sqrt{N}\log N\right)$
  - Pre and post-optimization: move of values across groups
  - Anytime meta-heuristic to improve solution
  - Criterion evaluated on discrete solutions
    - Numerical gap between evaluated solutions
    - No stopping threshold, iteration number or convergence rate to tune

Modeling with data dependent prior
The question of consistency

- Problem with modeling the finite data sample
  - The true real-valued joint-density distribution is not modeled
- What about prediction regarding other data samples?
- What about asymptotical consistency?

Consistency of the approach
Theorem 1

- Theorem 1: (EGC 2012)
  - The MODL evaluation criterion of a coclustering model *M* is asymptotically equal to *N* time the sum of the variable entropies minus mutual information between the grouped variables

$$c(M) = N\left(H\left(X\right)+H\left(Y\right)-I\left(X_M;X_M\right)\right)+O\left(\log N\right)$$

  - Entropy $H(X) = -\sum_{v\in X} p(v)\log p(v)$
  - Mutual information $I(X;Y) = \sum_{v\in X}\sum_{w\in Y} p(v,w)\log\frac{p(v,w)}{p(v)\,p(w)}$

- The approach tends to maximize the contrast between the variables
  - Mutual information between the grouped variables
    - Measures the mutual dependence between two variables, null in case of independence

Consistency of the approach
Theorem 2

- Theorem 2: (EGC 2012)
  - The MODL approach for selecting a coclustering model *M* asymptotically converges towards the true joint distribution of the two variables, and the criterion for the best model M_{Best} converges to *N* times the joint entropy of the variables.

$$\lim_{N\to\infty} \frac{c\left(M_{Best}\right)}{N} = H\left(X,Y\right)$$

- The approach is consistent
  - the model selected according the MODL approach converges towards that of the true joint density

Consistency of the approach
Corrollary

- The best model M_{Best} provides an estimator of the mutual information between the variables.

$$\lim_{N\to\infty} \frac{c(M_{\varnothing}) - c(M_{Best})}{N} = I(X;Y)$$

  - $I(X; Y)=0$ if and only if X and Y are independent

- The approach is robust
  - The null model (one single cell) is asymptotically selected in case of independence between the variables
  - Valid non asymptotically owing to the prior terms that penalize complex models
  - Confirmed experimentally on data samples of any size

Synthesis: analysis of the criterion

- Nul model (one single group per variable)
  - Each variable is coded independently

$$c(M_{\varnothing}) = N\,\mathrm{H}(X) + N\,\mathrm{H}(Y) + O(\log N)$$

Independence

- Intermediate model (a data grid)
  - Coding of the frequencies of cells of the data grid model
  - Plus coding of the data given the data grid model

Correlation

- Optimal model
  - The most probable
  - The minimum description length model
  - Approximation of the Kolmogorov complexity of the data
  - Allow to find the optimal granularity of the correlation model

Optimal

- Theorem
  - The MODL approach for coclustering of two variables asymptotically converges towards the true joint entropy of the variables

$$\lim_{m\to\infty} \frac{c(M_{Best})}{N} = H(X,Y) \qquad \lim_{N\to\infty} \frac{c(M_{\varnothing}) - c(M_{Best})}{N} = I(X;Y)$$

Consistency of the MODL approach

Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

What we expect: fast convergence rate
Density estimation: artificial dataset

- Circular random graph: no clusters
  - n=100 vertices equidistant on a unit circle

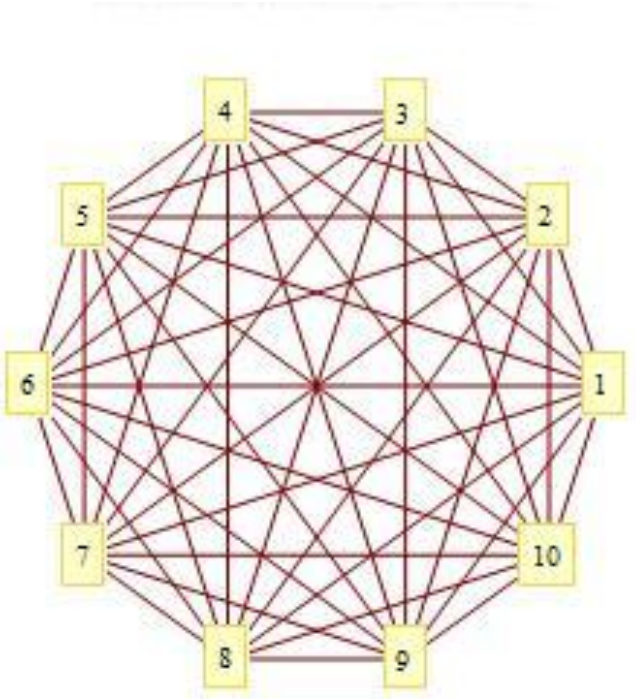
$$x_i = \cos\frac{2\pi i}{n},\ y_i = \sin\frac{2\pi i}{n}$$

  - Euclidian distance between vertices

$$d_{\substack{ij \\ i\neq j}} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2},\ d_{ii} = \frac{2}{n}$$

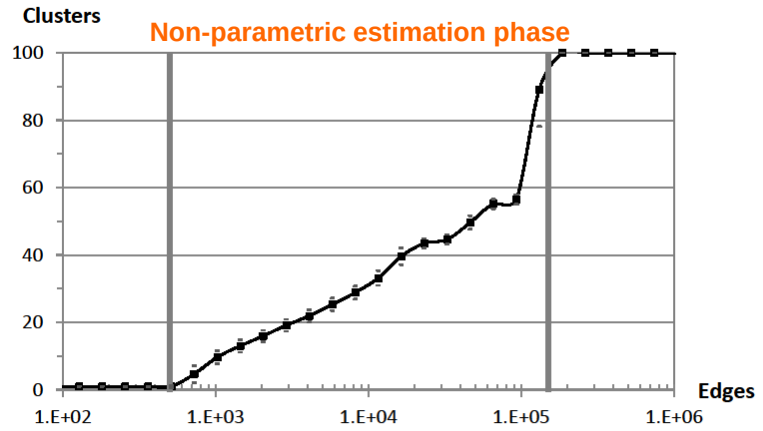
  - Probability of edges between two vertices in inverse proportion of their distance
    - problem: estimation of 10 000 edge probabilities

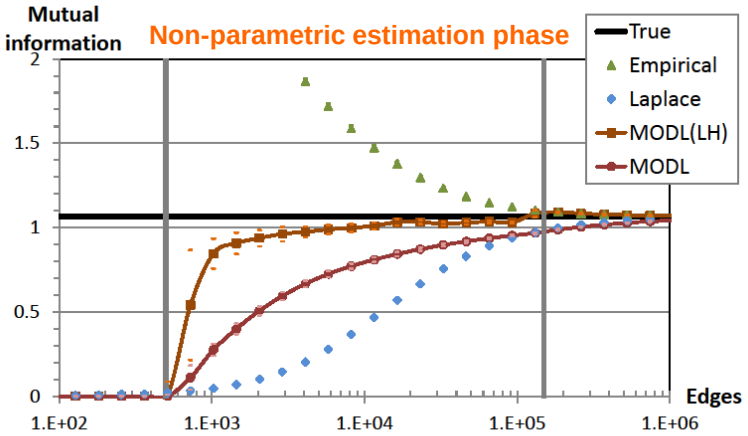
$$p_{ij} = \frac{\frac{1}{d_{ij}}}{\sum_{\mu,\gamma}\frac{1}{d_{\mu\gamma}}}$$



Convergence rate
Density estimation: results

- Experiment
  - Random samples of size from 100 to 10^6
  - Repeated 100 times
- Collected results
  - Cluster number
  - Estimated mutual information
- Three phases of convergence
  - Stability phase
    - Insufficient number of edges
    - One single cluster
  - Non-parametric estimation phase
    - Increasing number of clusters
    - Fast improvement of estimation quality
  - Parametric estimation phase
    - As many clusters as vertices
    - Classical improvement of estimation with sample size





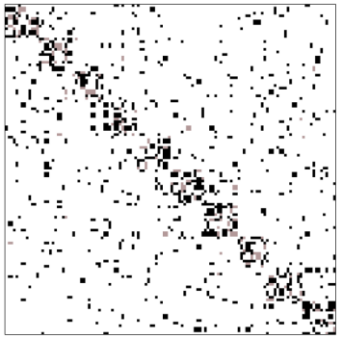
What we expect: discover patterns

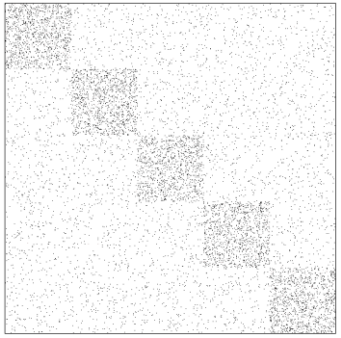
Pattern discovery: artificial datasets

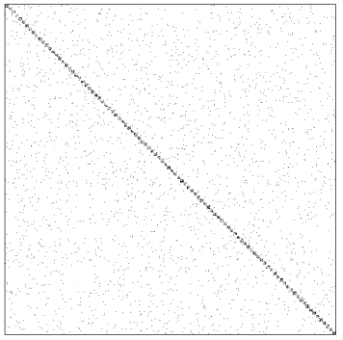
- Block-diagonal graphs: many clusters
  - Chose of equal size clusters
  - Pure clusters: 0% noise
  - Noisy clusters: 50% noise
  - Random graphs: 100% noise

Vertices	Clusters
10	2
100	10
1000	5
1000	100
10000	200

  - Problem: reliably identify the pattern

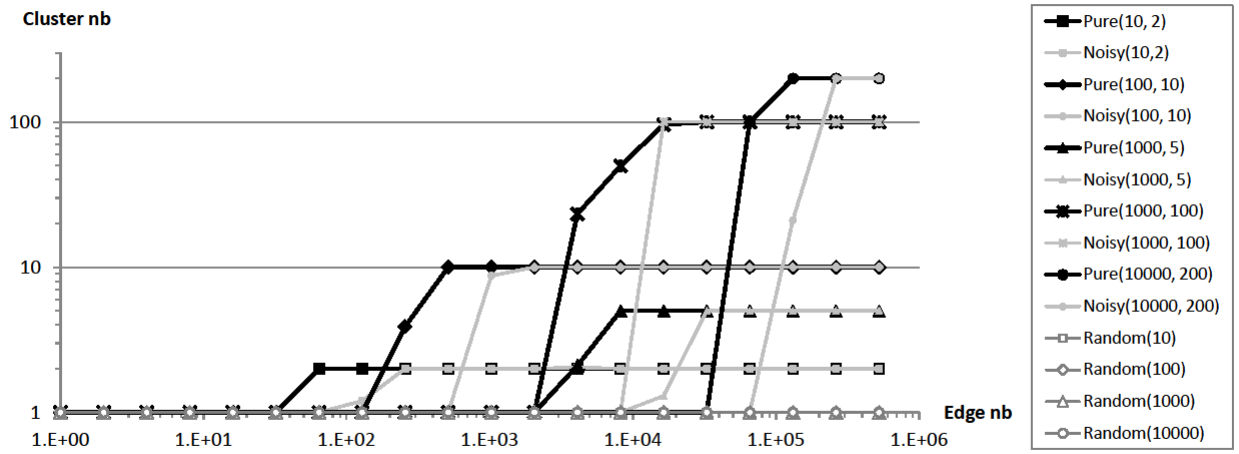






Convergence rate
Pattern discovery: results

- Experiment
  - Random samples of size from 1 to 10^6, repeated 10 times
  - Collected results: cluster number
- Fast convergence
  - Fast transition between stability phase and parametric estimation phase
  - Noisy patterns need four times more instances
  - Never produce spurious clusters
  - Random graphs always summarized with one single cluster



Schedule

- Coclustering and its Applications
- Survey of Coclustering Methods
- Data Grid Models for Coclustering
- Consistency of the Approach
- Evaluation
- Conclusion

Conclusion

- MODL approach for selection of coclustering models
  - Bayesian approach of model selection in discrete model space
  - Data dependent model space and prior!
  - Important result: consistency of the approach
- Extensive empirical validation
  - Both on artificial and real world datasets
  - Download as a shareware: *www.khiops.com*
- Numerous applications of coclustering
  - Clustering of texts, graphs, curves
  - Web mining (content, usage, structure)
  - Market basket analysis
  - …
- Open theoretical questions
  - Consistency in case of numerical variables
  - Consistency in the multivariate case (K>2)
  - Convergence rate

Thank you for your attention!

Coclustering applications using data grid models

Alexis Bondu, Marc Boullé, Dominique Gay

January, 29, 2013

Schedule

- **Spatial data**
  - → Geographical distribution of industries
  - → Air traffic
- **Spatio-temporal data**
  - → Rental bike service (London)
  - → Phone call log
- **Time series**
  - → Individual electricity consumption
  - → Rainfall
- **Text mining**
  - → News Groups
  - → Audio transcription : phone counseling to customers
- **WEB mining**
  - → Spam *(Web Structure Mining)*
  - → Log *(Web Usage Mining)*

Spatial data

A – Geographical distribution of industries (Paris)

734 NAF codes

NAF Code	Geographical Units

Companies (605 639)

5146 streets of Paris

Spatial data

A – Geographical distribution of industries (Paris)

- Optimal model :
  - → 50 groups of NAF code
  - → 234 groups of streets

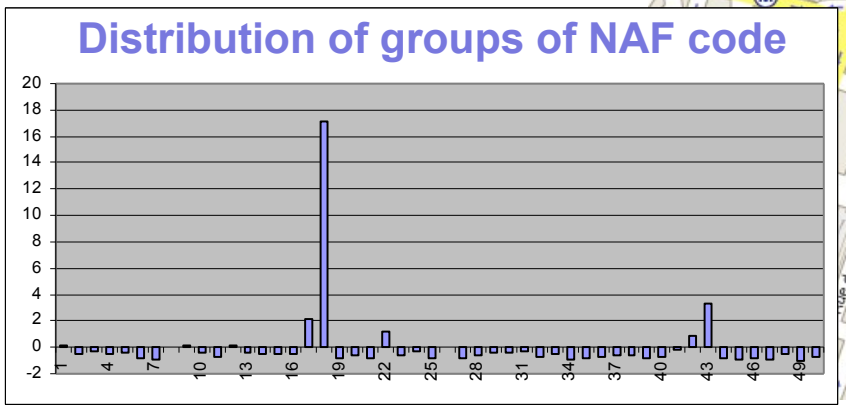
Group « Art and entertainment »:
- Performing arts
- Other creative arts
- Artistic creation within visual arts

Spatial data

A – Geographical distribution of industries (Paris)

- Optimal model :
  - → 50 groups of NAF code
  - → 234 groups of streets

Group of streets « le sentier »:

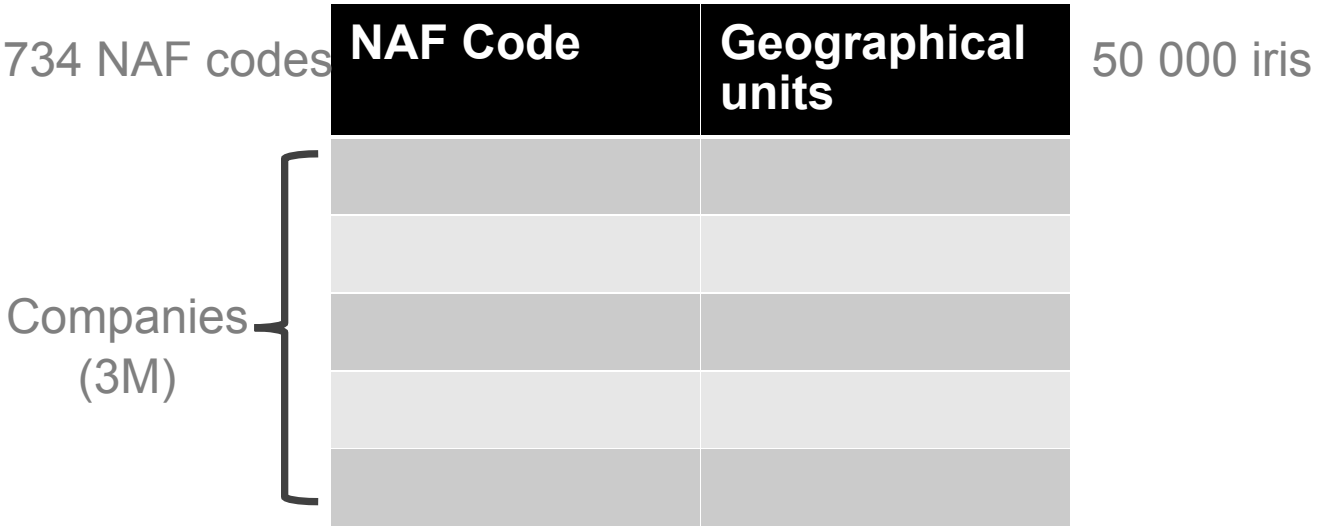


(Manufacturing and wholesale textile, Industrial Zone textiles & jewelry, Retail Sales, wholesale trade …)

(21 streets)

Spatial data

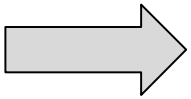
B – Geographical distribution of industries (France)



Spatial data

B – Geographical distribution of industries (France)

- Optimal model :
  - → 82 groups of NAF code
  - → 323 groups of Iris



Folding of the optimal model : 18 selected groups of NAF code

- ✓ Secteur primaire
- ✓ Centre-ville
- ✓ Art et graphismes
- ✓ Commerce de Gros
- ✓ Finance
- ✓ Activités juridiques
- ✓ Informatique
- ✓ Services liés à la santé
- ✓ Industrie, autres ventes en gros et transports liés
- ✓ Location immobilière, restauration
- ✓ Traditionnelle, hôtels et commerces de boisson
- ✓ Production culturelle, marketing et services aux entreprises
- ✓ Production, transport et commerces en lien avec le milieu énergétique
- ✓ Textile
- ✓ Tourisme mer et montagne
- ✓ Tourisme vert
- ✓ Travaux immobiliers
- ✓ Recherche et développement, ingénierie
- ✓ Divers

Spatial data

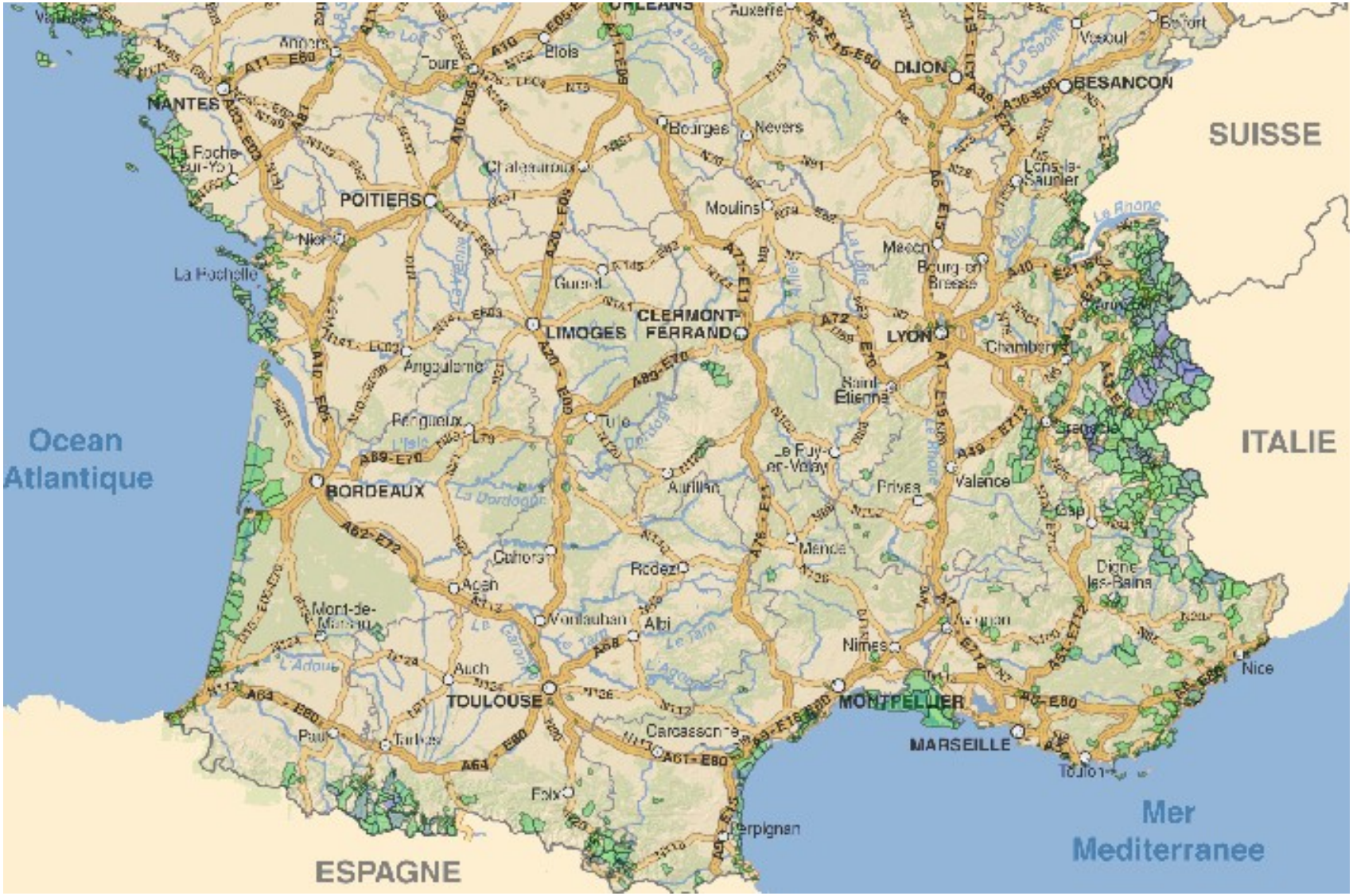
B – Geographical distribution of industries (France)

- Group of NAF code « Sea and mountain tourism » :

Code NAF - label	Fréquence
6820A - Location de logements	14048
8551Z - Enseignement de disciplines sportives et d'activités de loisirs	5810
0311Z - Pêche en mer	1264
9319Z - Autres activités liées au sport	722
0321Z - Aquaculture en mer	533
3315Z - Réparation et maintenance navale	443
7721Z - Location et location-bail d'articles de loisirs et de sport	349
4638A - Commerce de gros (commerce interentreprises) de poissons, crustacés et mollusques	263
5222Z - Services auxiliaires des transports par eau	150
5010Z - Transports maritimes et côtiers de passagers	106
3012Z - Construction de bateaux de plaisance	90
1020Z - Transformation et conservation de poisson, de crustacés et de mollusques	79
4939C - Téléphériques et remontées mécaniques	79
4291Z - Construction d'ouvrages maritimes et fluviaux	65
0893Z - Production de sel	63
3011Z - Construction de navires et de structures flottantes	42
1092Z - Fabrication d'aliments pour animaux de compagnie	18

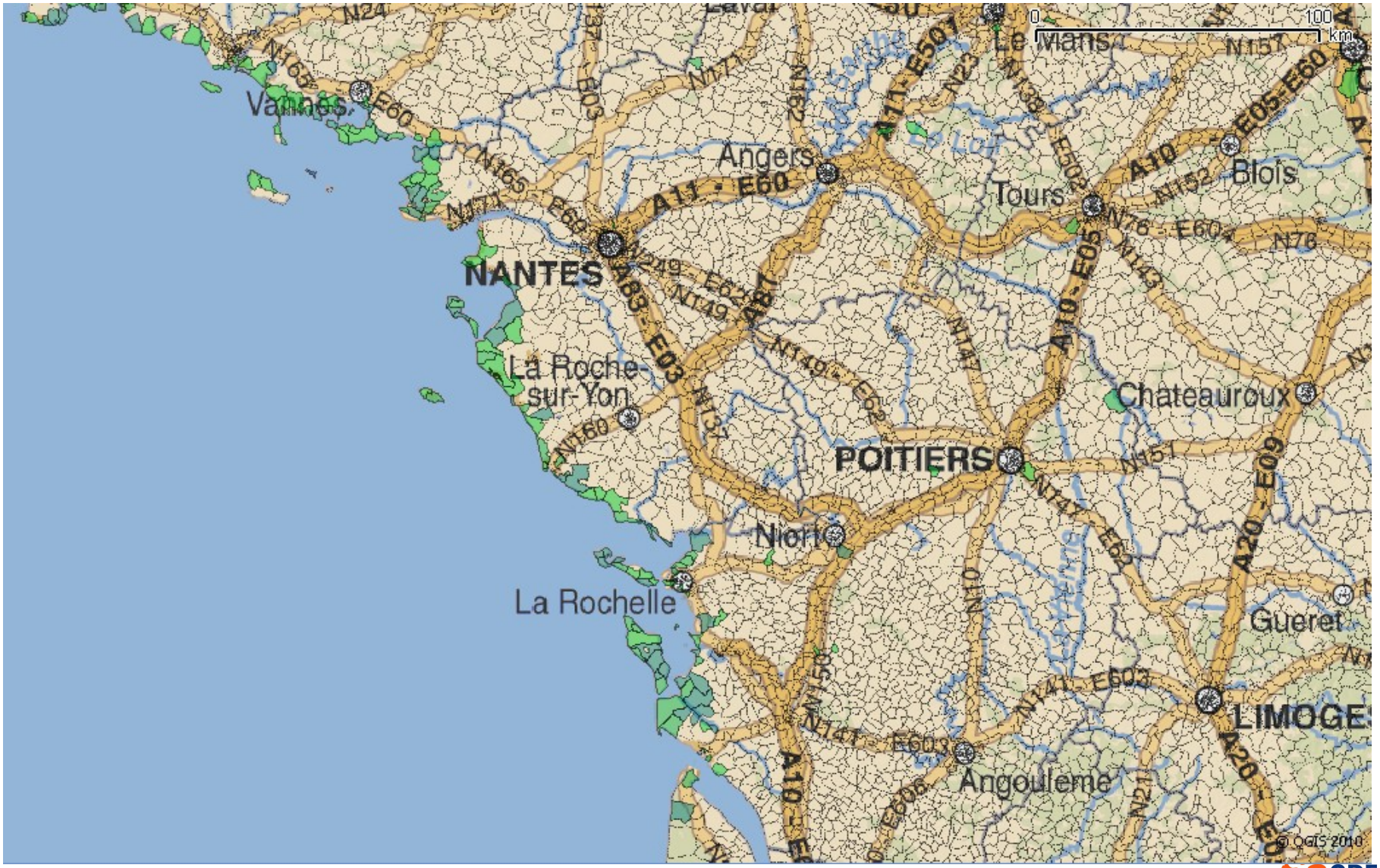
Spatial data

Iris with high probability of group of NAF code « Sea and mountain tourism » :



Spatial data

Iris with high probability of group of NAF code « Sea and mountain tourism » :



Spatio-temporal data :

Rental bike service (Londres)

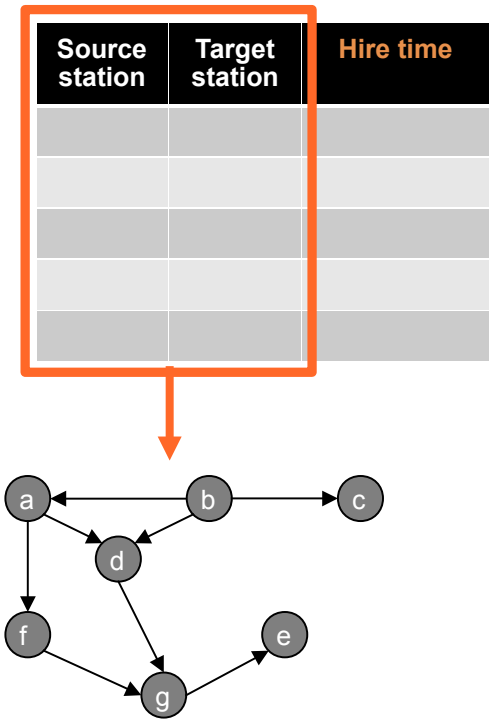
488 stations | Minute precision

Source station	Target station	Hire time

Journeys (4,8 M)

Spatio-temporal data :

Rental bike service (Londres)

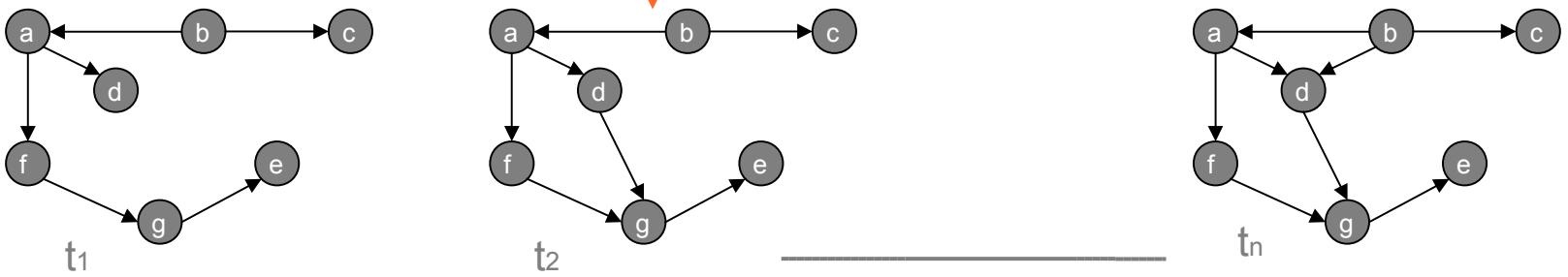


Description of an directed graph

Spatio-temporal data :

Rental bike service (Londres)

Source station	Target station	Hire time



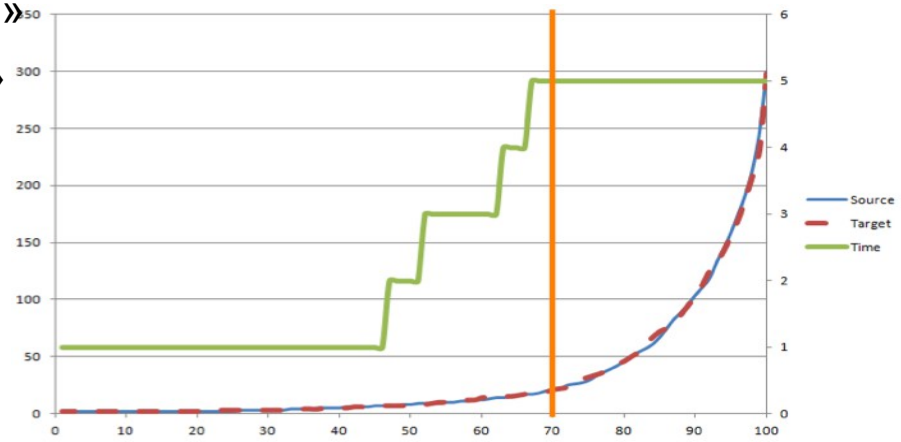
Description of a temporal directed graph

Spatio-temporal data :

Rental bike service (Londres)

- Optimal model :
  - → 296 groups of « source stations »
  - → 281 groups of « target stations »
  - → 5 time slots

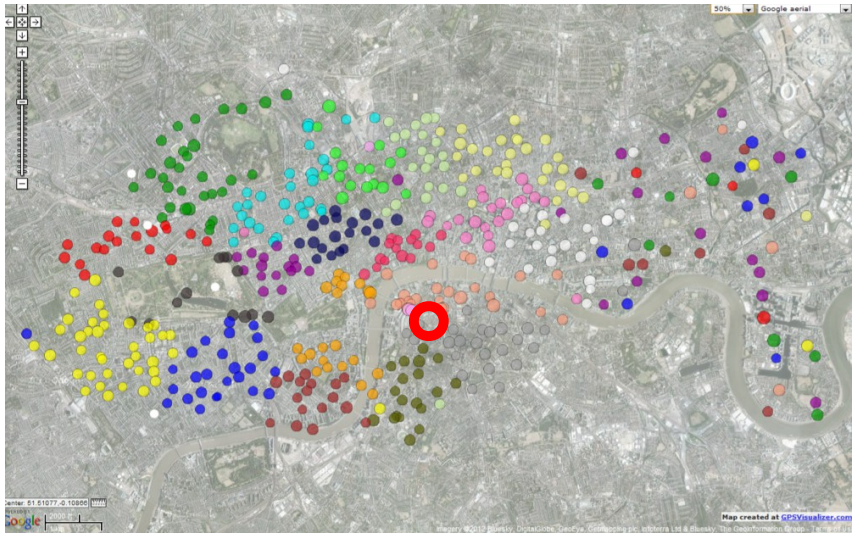
- Selected model : retains 70% of the information
  - → 20 groups of « source stations »
  - → 20 groups of « target stations »
  - → 5 time slots



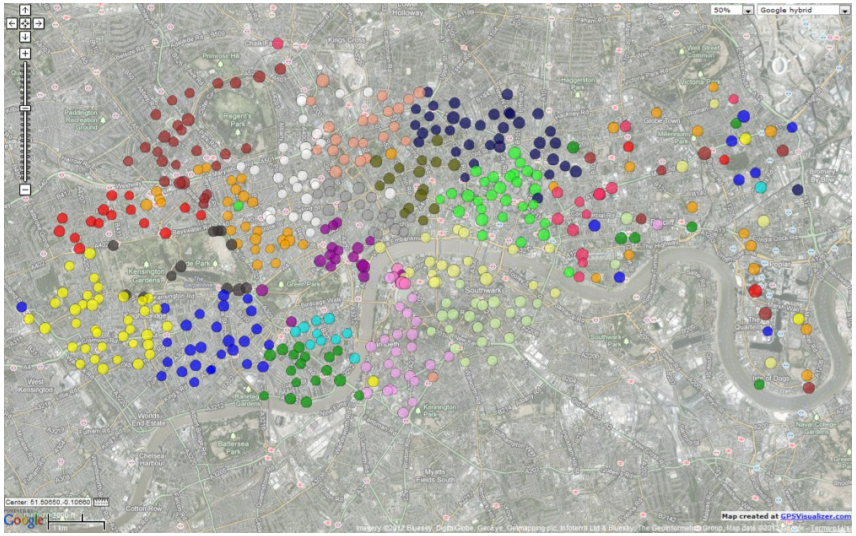
Spatio-temporal data :

Rental bike service (Londres)

Groups of « source stations »



Groups of « target stations »

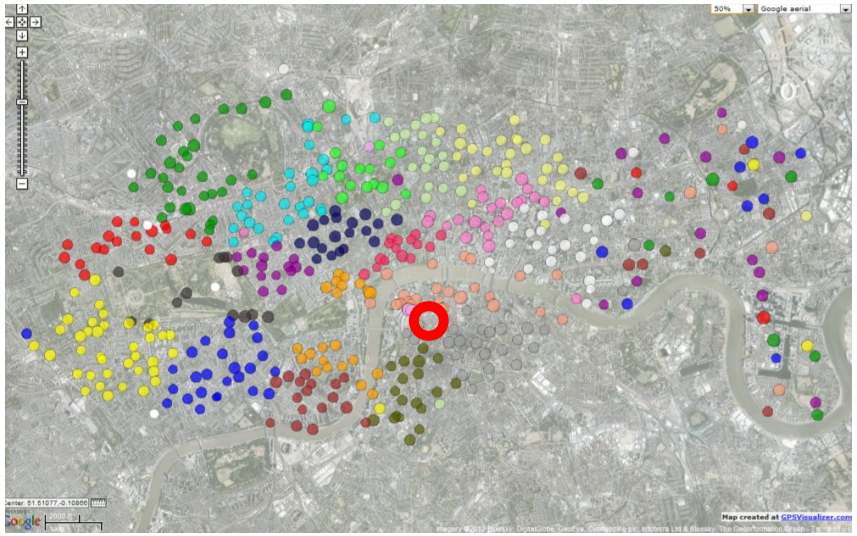


Where go the bikes which start from Waterloo before 7:00 am ?

Spatio-temporal data :

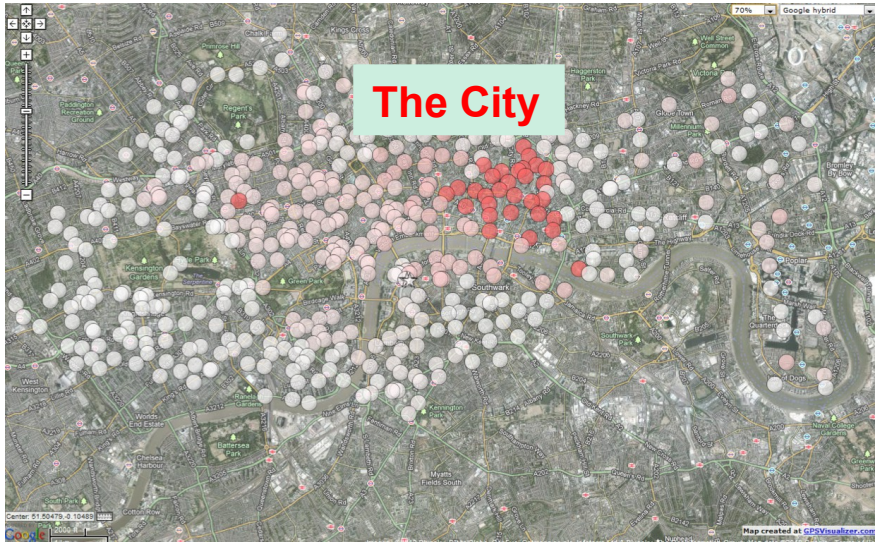
Rental bike service (Londres)

Groups of « source stations »



Where go the bikes which start from Waterloo before 7:00 am ?

Mutual information between clusters, on a particular time slot



Time series :

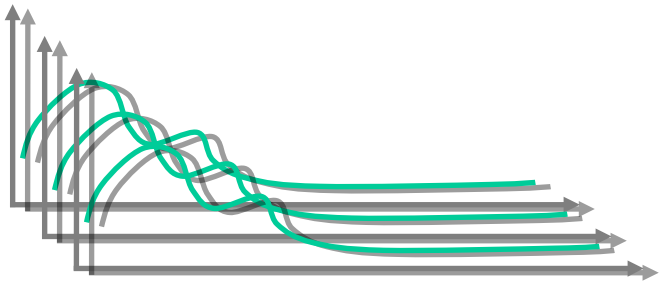
Individual electricity consumption

Exploited dataset :

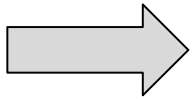
- A single household
- Daily consumption, sampling rate 10” *(144 measures/day)*
- 349 days

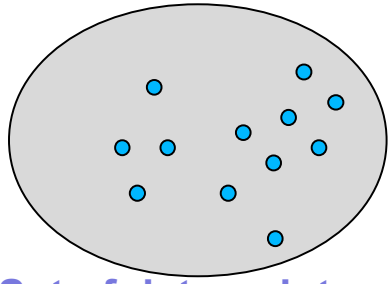
Time series :

Individual electricity consumption



Collection of time series





Set of data points

349 series

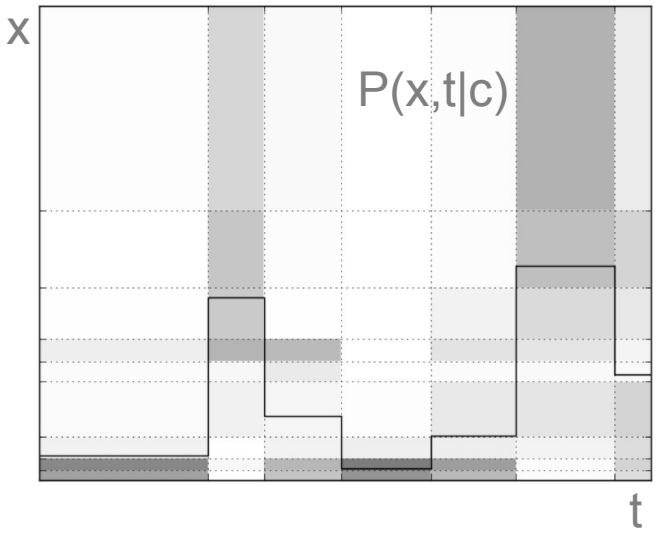
Series Id *(Day)*	Timestamp	Value

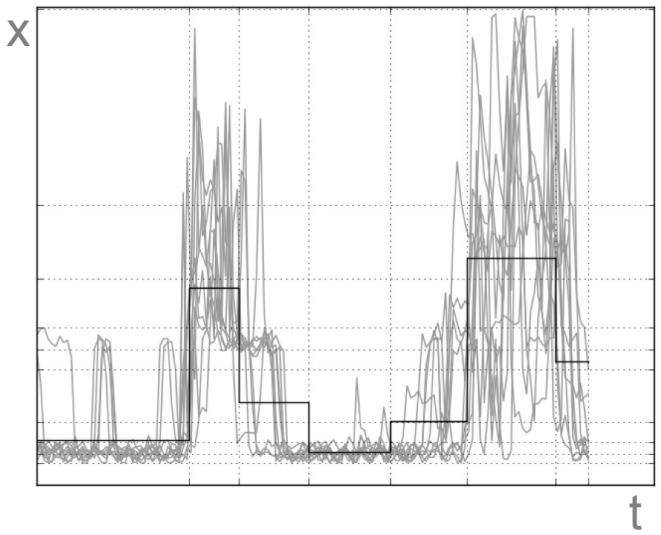
Data points (50 000)

Time series :

Individual electricity consumption

- Optimal model
  - → 57 groups of days
  - → 7 time slots
  - → 10 intervals of consumption values



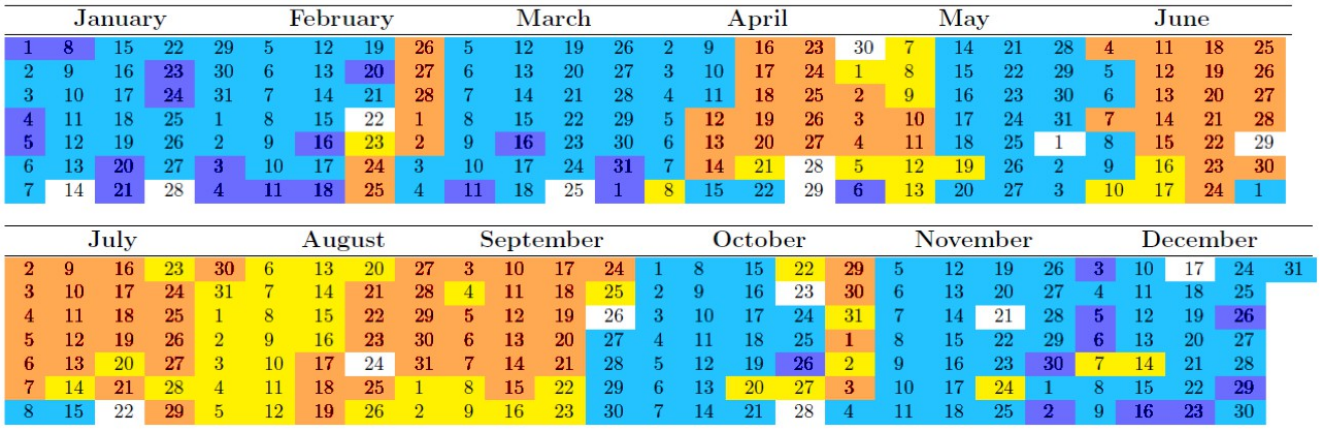


Time series :

Individual electricity consumption

- Selected Model : retains 50% of the information
  - → 4 groups of days
  - → 4 time slots
  - → 4 Intervals of consumption values

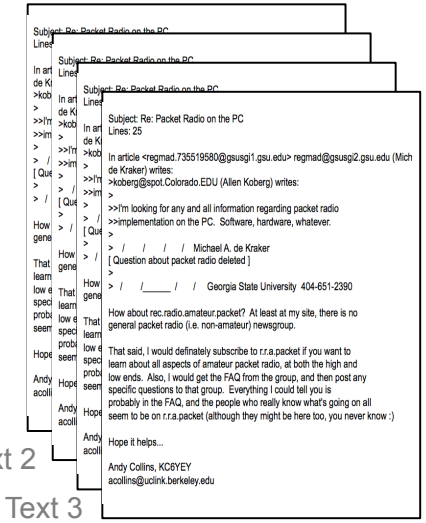
Representation of the groups of days on a calendar



January									February									March									April									May									June								
1	8	15	22	29	5	12	19	26	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	3	10	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24	31
2	9	16	23	30	6	13	20	27	6	13	20	27	3	10	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	4	11	18	25	1
3	10	17	24	31	7	14	21	28	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	5	12	19	26	2
4	11	18	25	1	8	15	22	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	6	13	20	27	3
5	12	19	26	2	9	16	23	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	7	14	21	28	4
6	13	20	27	3	10	17	24	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	31	8	15	22	29	5
7	14	21	28	4	11	18	25	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	1	8	15	22	29	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	1	9	16	23	30	6

July									August									September									October									November									December								
2	9	16	23	30	6	13	20	27	3	10	17	24	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	4	11	18	25					
3	10	17	24	31	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	5	12	19	26					
4	11	18	25	1	8	15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	6	13	20	27					
5	12	19	26	2	9	16	23	30	6	13	20	27	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	7	14	21	28					
6	13	20	27	3	10	17	24	31	7	14	21	28	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	31	8	15	22	29					
7	14	21	28	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	1	8	15	22	29	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	1	9	16	23	30					
8	15	22	29	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	3	10	17	24	31	8	15	22	29	5	12	19	26	2	10	17	24	31					

Text mining : *News groups*



Words in texts (2M)

20 000 texts 20 000 words

Text ID	Word

Challenge IJCNN 2007 : news groups dataset (Mitchell 99)

A supervised problem :

Topics of classe 1

atheism, graphics, ms-windows , ibm.pc.hardware, mac.hardware, windows.x, forsale, autos, motorcycles, baseball, hockey, crypt, electronics, med, space

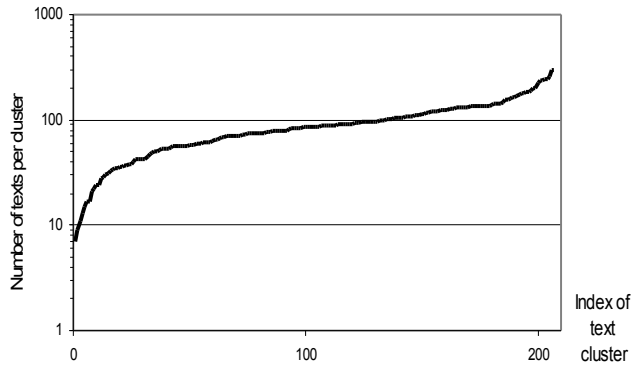
Topics of classe 2

religion.christian, politics.guns, politics.mideast, politics.misc, religion.misc

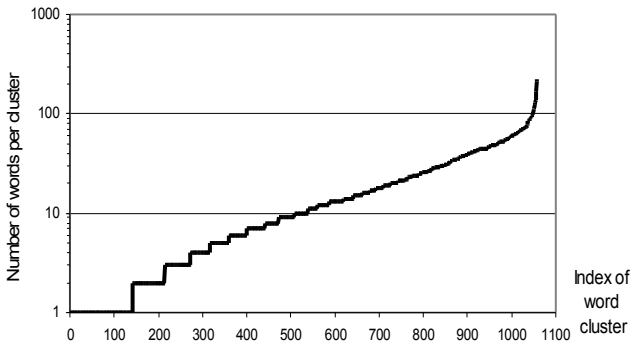
Text mining : *News groups*

- Optimal model :
  - → 207 groups of texts
  - → 1058 groups of words

Size of groups of texts *(balanced)*



Size of groups of words *(unbalanced)*



Text mining : *News groups*

Groupe of words : the most correlated with classes

- hockey, playoff, nhl, penguin, devils, pens, leafs, bruins, islande, goalie, mario, puck,...
- team, season, league, fans, teams, rangers, detroit, montrea, wins, scored, coach,...
- clipper, encrypt, nsa, escrow, pgp, crypto, wiretap, privacy, cryptog, denning,...
- dod, bike, motorcy, ride, riding, bikes, ama, rider, helmet, yamaha, harley, moto,...
- basebal, sox, jays, giants, mets, phillie, indians, cubs, yankees, stadium, cardina,...
- bible, scriptu, teachin, biblica, passage, theolog, prophet, spiritu, testame, revelat,...
- christi, beliefs, loving, rejecti, obedien, desires, buddhis, deity, strive, healed,...
- windows, dos, apps, exe, novell, ini, borland, ver, lan, desqvie, tsr, workgro, sdk,...
- pitcher, braves, pitch, pitchin, hitter, inning, pitched, pitches, innings, catcher,...
- car, cars, engine, auto, automob, mileage, autos, cactus, pickup, alarm, sunroof,...

Text mining : *News groups*

Groupe of words : the less correlated with classes

- book, books, learnin, deals, booksto, encyclo, titled, songs, helper
- cause, caused, causes, occur, occurs, causing, persist, excessi, occurin
- importa, extreme, careful, essenti, somewha, adequat
- morning, yesterd, sunday, friday, tuesday, saturda, monday, wednesd, thursda,...
- receive, sent, placed, returne, receivi, sends, resume

Text mining : *News groups*

Supervised evaluation :

1. Coclustering (id_text, word)
   Train set + Validation set + Test set : 19466 texts

1. Learning of a classifier
   Train set + Validation set : 1829 texts

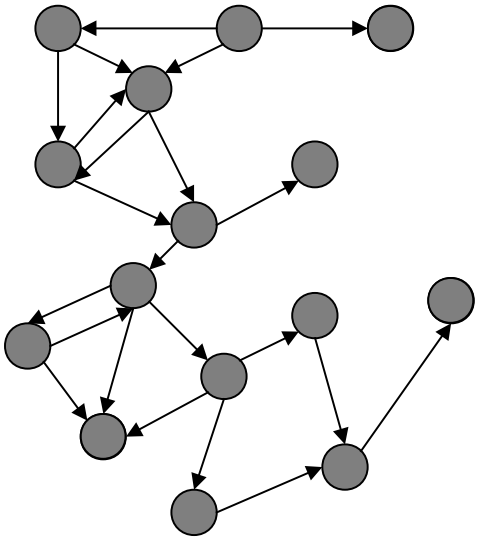
Results

The best classifier of the challenge : 4.6% *(balanced error rate)*
Coclustering MODL : 3.70% *(balanced error rate)*

✓ Better than the winner of the IJCNN 2007 challenge

Web mining

A – Spam *(web structure mining)*



Links between hosts
(500 000)

9 000 Hosts	9 000 Hosts
Source Host	**Target Host**

Challenge ECML/PKDD 2007: detection of web spam

A supervised problem :

Classe 1 : normal web host

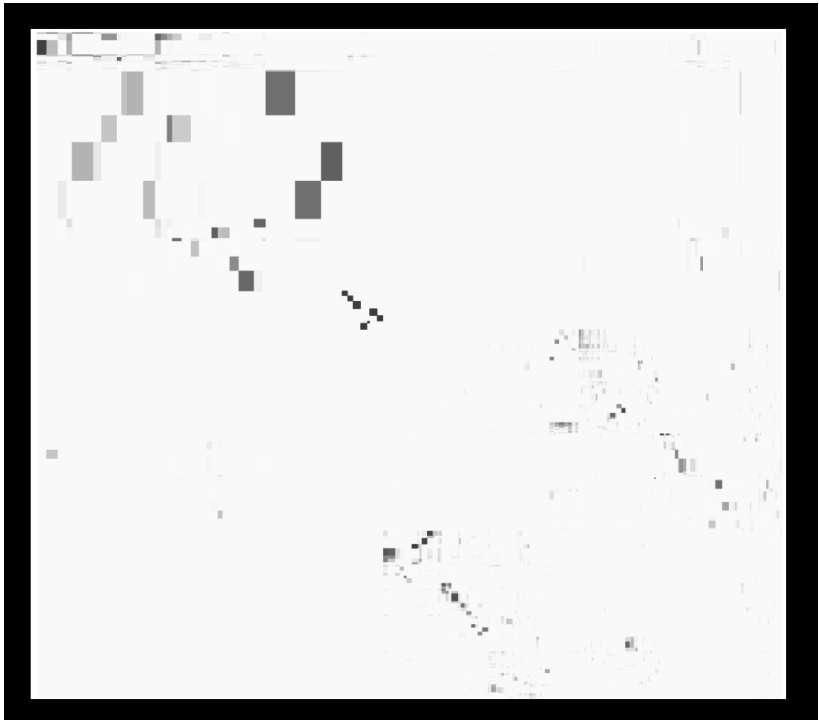
Classe 2 : spam web host

Web mining

A – Spam *(web structure mining)*

- Optimal model :
  - → 167 groups of source hosts
  - → 219 groups of target hosts

Mutual information between clusters



Web mining

A – Spam *(web structure mining)*

Supervised evaluation :

1. Coclustering *(source_host, target_host)*
   Train set + Test set

1. Folding of the optimal model *(200, 100, 50 clusters source + target)*

1. Learning of a classifier exploiting clusters *(Selective Naive Bayes)*
   Train set

Results

The best classifier of the challenge : 0.952 *(AUC)*
Coclustering MODL + SNB : 0.951 *(AUC)*

✓ Equivalent to the best classifier of the ECML/PKDD challenge

Web mining

B – Web log *(web usage mining)*

navigation from Orange search engine

270 000 cookies 40 000 Web sites

Cookie	Web site

Initial matrix with 10 billions cells

Web log
cookie*web site
(6 600 000)

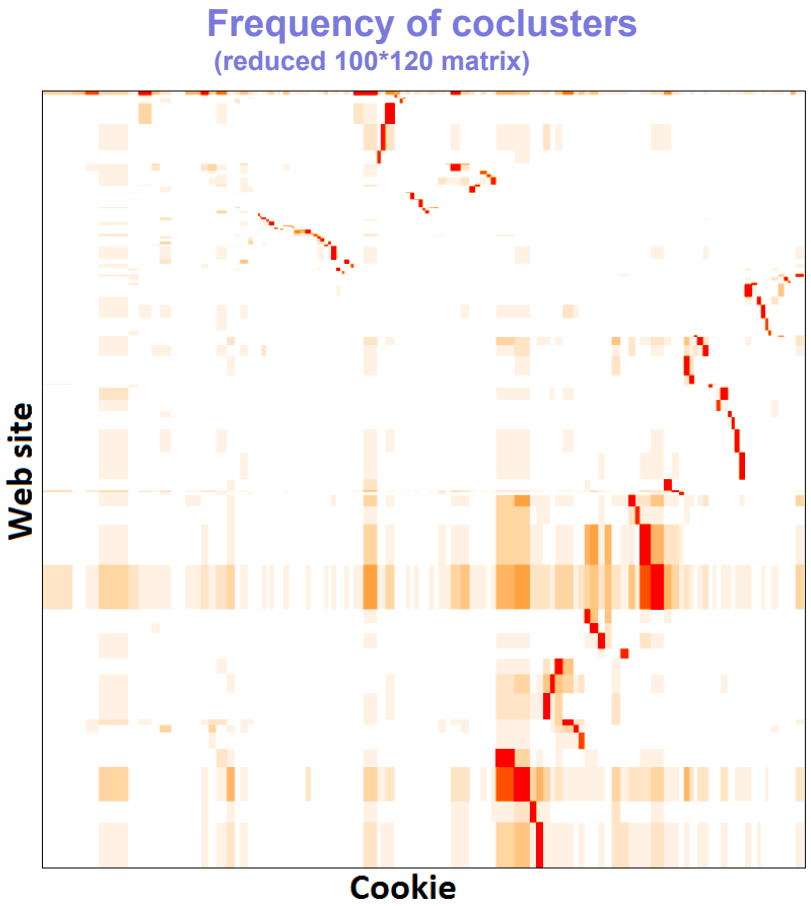
Exploratory analysis :

Thematic clustering of web sites
Behavioral clustering of cookies

Web mining

B – Web log *(web usage mining)*

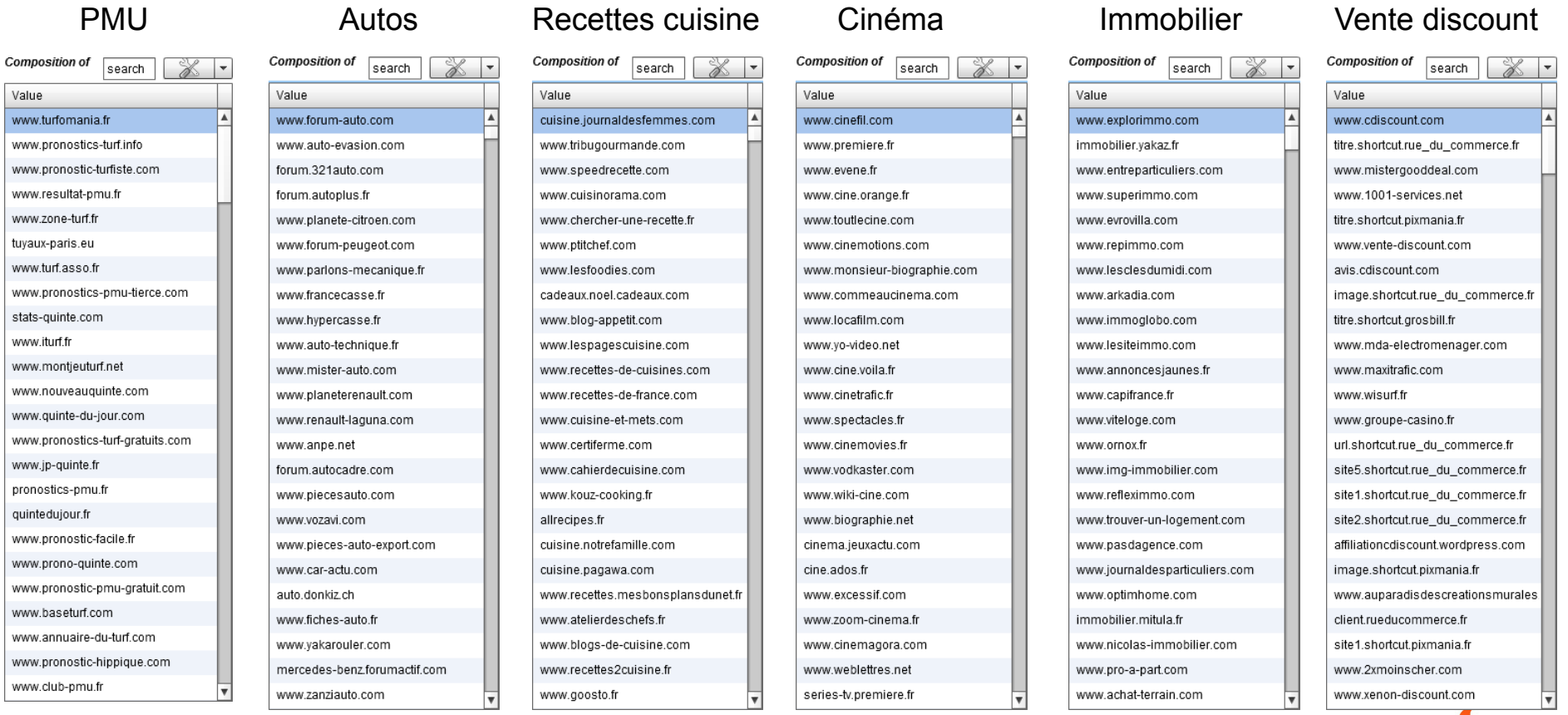
- Optimal model :
  - → 427 groups of web sites
    (thematic clustering)
  - → 1612 groups of cookies
    (behavioral clustering)



Web mining

B – Web log *(web usage mining)*

Exploratory analysis : examples of clusters of web sites
Automatically identified from web log analysis

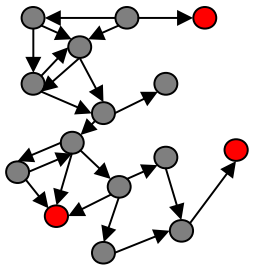


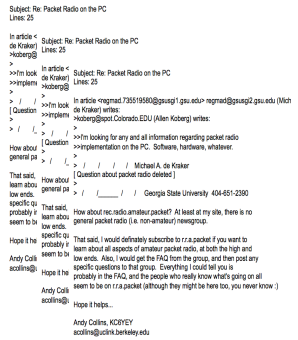
Conclusion

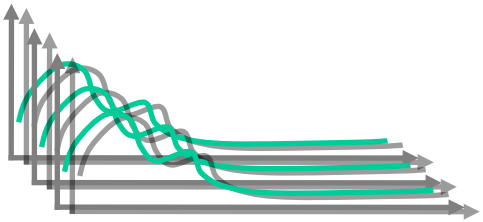
The MODL coclustering can be applied on a large range of :

Data types

- Graphs
- Temporal graphs
- Text datasets
- Time series
- Tabular datasets

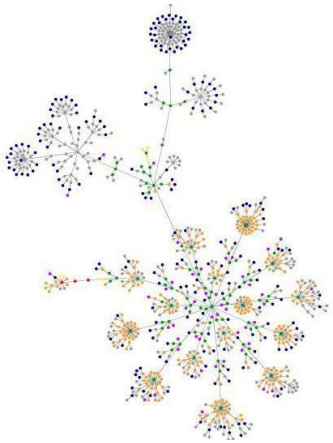






Application areas

- Web content mining
- Web structure mining
- Web usage mining
- Market basket analysis
- Recommender systems
- Geomarketing



Biblio

- **Spatial data**

  → Geographical distribution of industries

A. Bondu, M. Boullé, A. Peradotto, *« Représentation géographique du tissu industriel : une application de l'approche MODL »*, CAP 2010

  → Air traffic

M. Boullé. *« Nonparametric Edge Density Estimation in Large Graphs »,* Research Report France Telecom R&D, No FT/RD/TECH/11/02/13, 2011.

- **Spatio-temporal data**

  → Rental bike service (London)

R. Guigourès, M. Boullé, F. Rossi. *« Triclustering pour la détection de structures temporelles dans les graphes »* MARAMI 2012

  → Phone call log

R. Guigourès, M. Boullé, F. Rossi . *« Étude des corrélations spatio-temporelles des appels mobiles en France »*, EGC 2013

- **Time series**

  → Individual electricity consumption

M. Boullé. *« Functional data clustering via piecewise constant nonparametric density estimation »*, Pattern Recognition, 45(12):4389-4401, 2012.

Biblio

- **Text mining**

  → Opinions on films

  M. Boullé. *« Data grid models for preparation and modeling in supervised learning »*, In Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1, I. Guyon, G. Cawley, G. Dror, A. Saffari (eds.), pp. 99-130, Microtome Publishing, 2011.

- **WEB mining**

  → Spam *(Web Structure Mining)*

  M. Boullé. *« Nonparametric Edge Density Estimation in Large Graphs »,* Research Report France Telecom R&D, No FT/RD/TECH/11/02/13, 2011.

  → Log *(Web Usage Mining)*

  A. Beck. *« Coclustering Content and Usage Data to Improve Customer Knowledge Discovery from Web Logs», Submitted WWW 2013*.

Data grid models for supervised learning

Alexis Bondu, Marc Boullé, Dominique Gay

January, 29, 2013

Schedule

- **Live Demo of “Khiops” software**
  - → Tutorial
  - → Demo

- **Change Detection in Data Streams**
  - → A new supervised approach
  - → An alternative approach
  - → Comparative experiments
  - → Conclusion

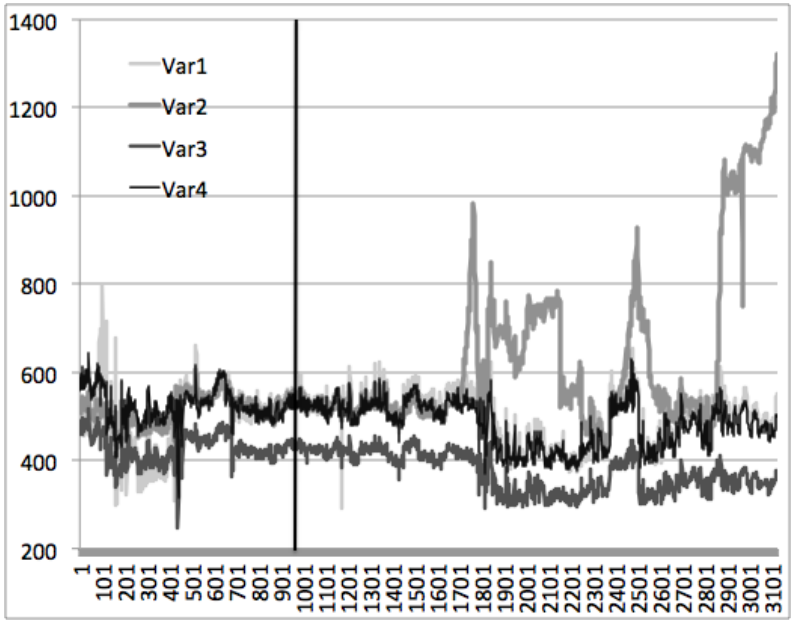
Schedule

- **Live Demo of “Khiops” software**
  - → Tutorial
  - → Demo

- **Change Detection in Data Streams**
  - → A new supervised approach
  - → An alternative approach
  - → Comparative experiments
  - → Conclusion

Change Detection in Data Streams

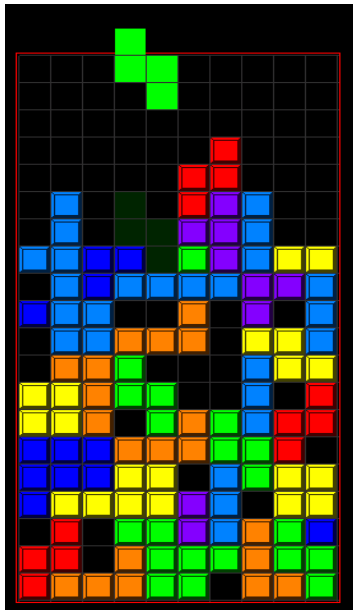
- Objective :
  - Change detection
  - Diagnostic



Towards an automated monitoring system

Change Detection in Data Streams

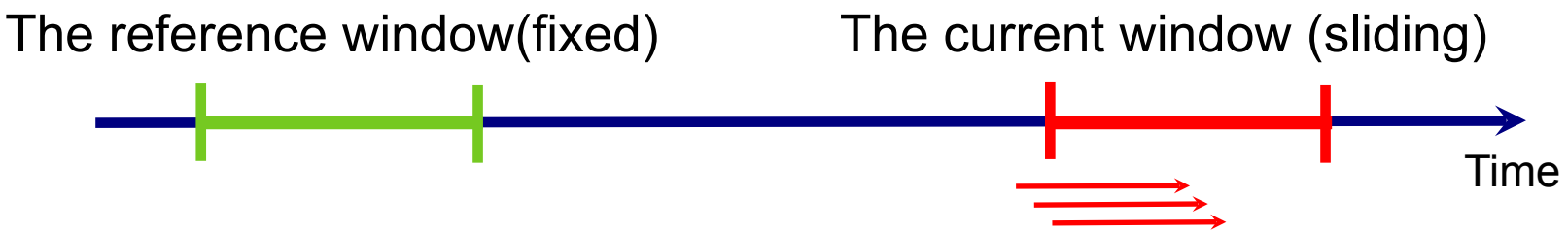
The paradigm of data streams



- Tuples are structured
- Their flow rate is not controlled
- Their order is not controlled

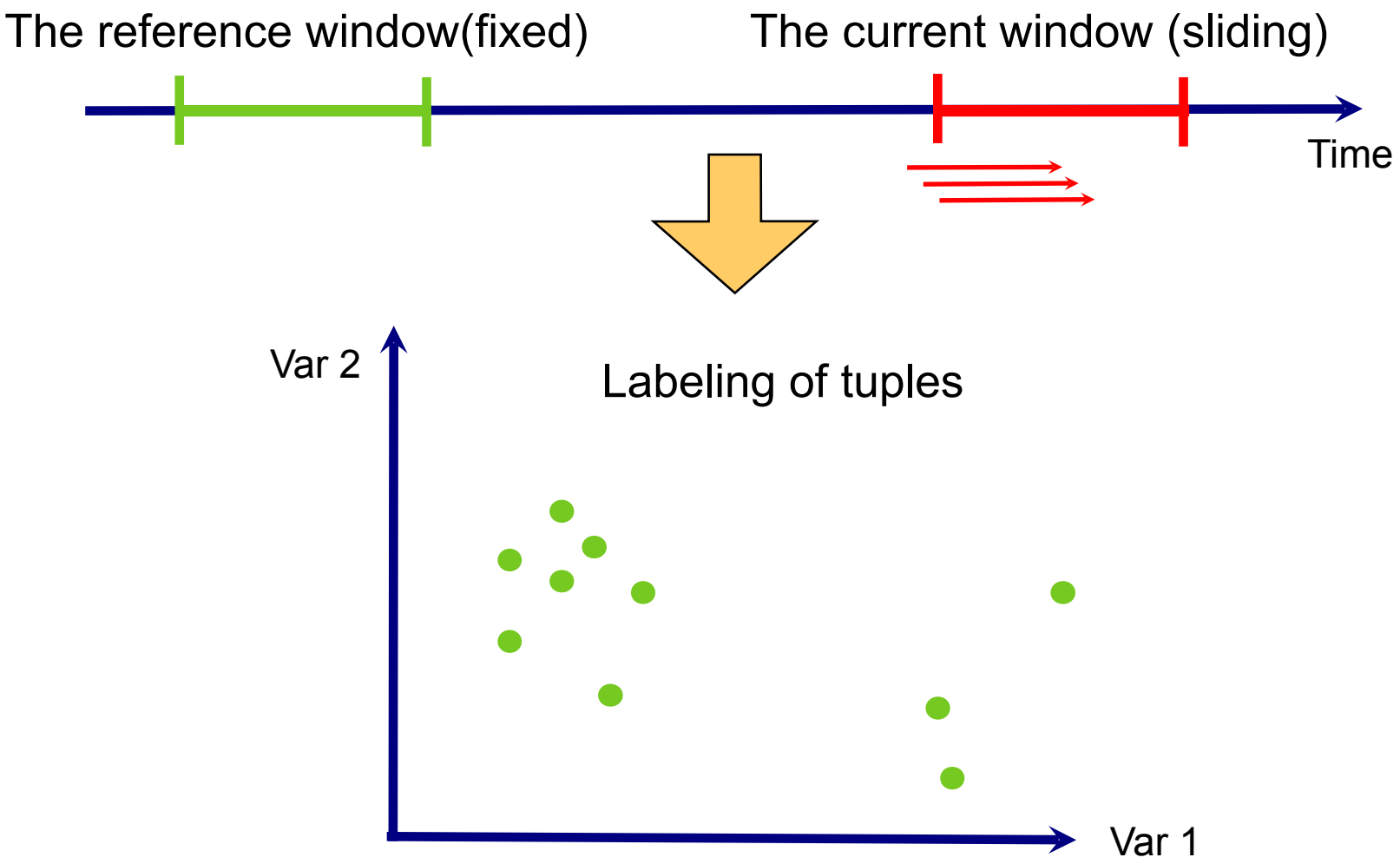
A new Supervised Approach

Let's consider a supervised problem ....



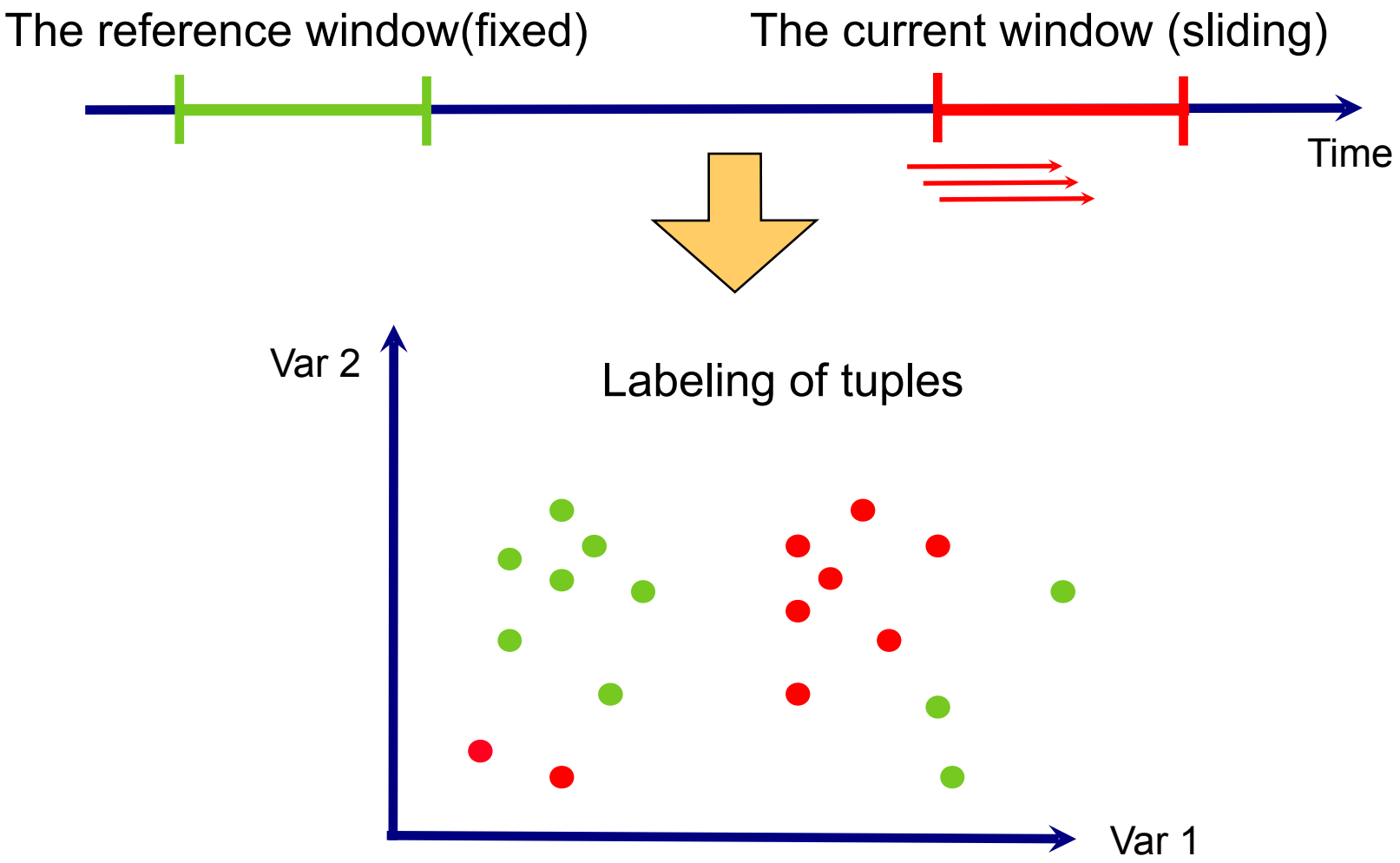
A new Supervised Approach

Let's consider a supervised problem ....



A new Supervised Approach

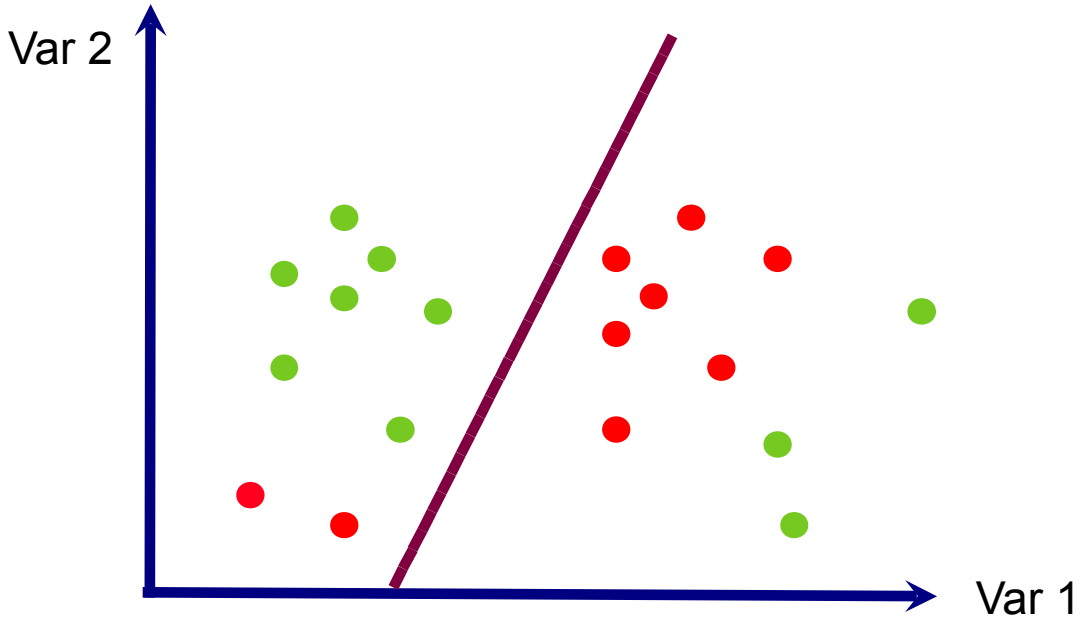
Let's consider a supervised problem ....



A new Supervised Approach

The insight

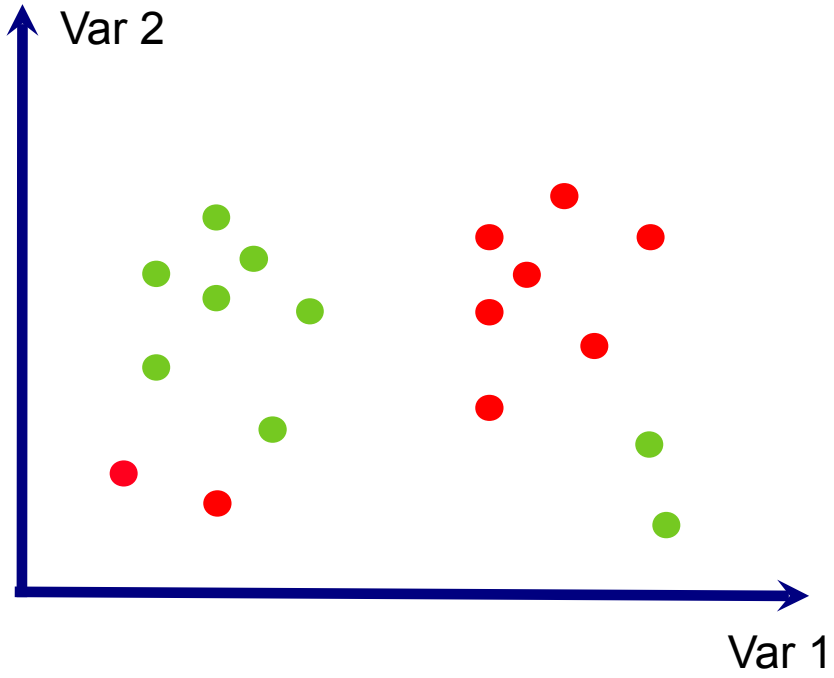
If classes can be discriminated, that means a change in the distribution of tuples has occurred.



A new Supervised Approach

A simplification .... not so drastic!

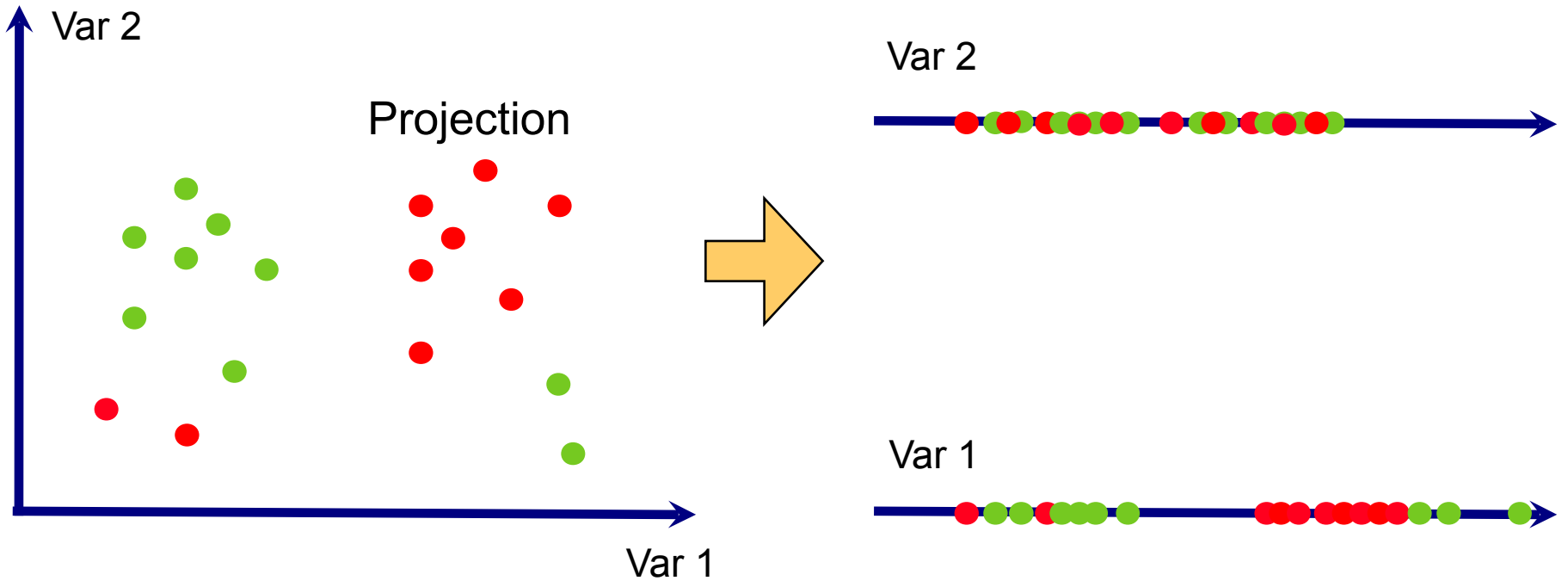
The tuples are projected on each variable, and several univariate classifiers are trained.



A new Supervised Approach

A simplification .... not so drastic!

The tuples are projected on each variable, and several univariate classifiers are trained.



A new Supervised Approach

The choice of the classifier is critical :

- The learning approach must be regularized :
  - avoid over-fitting
    - exploit all available tuples to learn (without evaluation set)

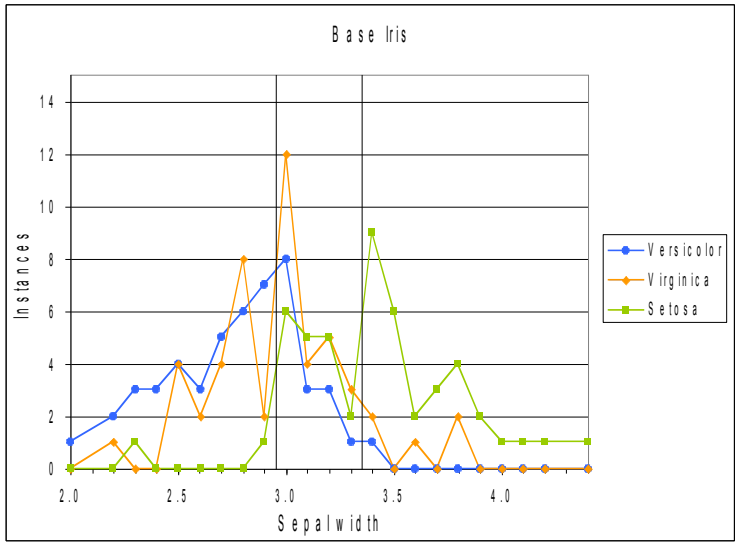
- The learning approach must be parameter-free

- The predictive model must estimate the conditional density
  - estimate the distribution of classes
  - evaluate the “gap” between these distributions

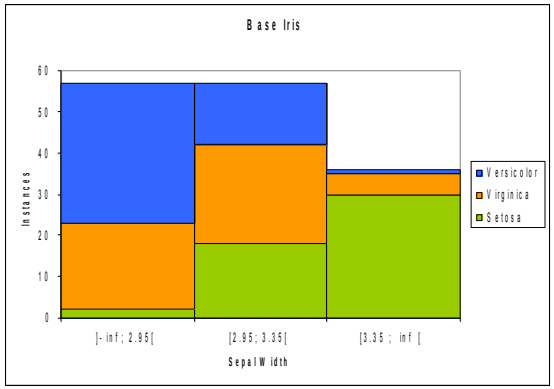
A new Supervised Approach

Our choice : The discretization method MODL

- Definition : cutting out the numerical domain of a variable into intervals.
- Objective : describe the conditional distribution of data.



Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Sepalwidth	]- inf ; 2.95 [	[2.95 ; 3.35 [	[3.35 ; inf [



What is the best model?

A new Supervised Approach

Change detection

Coding length (Shannon)

$$\mathcal{C}_k(M) = -\log[P(M|D_k)]$$

Compression gain (Boullé 2009)

$$\mathcal{G}ain_k(M) = 1 - C_k(M)/C_k(M0)$$

Coding length of the null model

Coding length of the current model

A new Supervised Approach

Change detection

Coding length (Shannon)

$$\mathcal{C}_k(M) = -\log[P(M|D_k)]$$

Compression gain (Boullé 2009)

$$\mathcal{G}ain_k(M) = 1 - C_k(M)/C_k(M0)$$

Quantifying the change

$$\mathcal{C}hange = \frac{1}{K} \sum_{k \in [1,K]} \mathcal{G}ain_k(\mathcal{M}ap)$$

The most probable model

A new Supervised Approach

Change detection

Coding length (Shannon)

$$\mathcal{C}_k(M) = -\log[P(M|D_k)]$$

Compression gain (Boullé 2009)

$$\mathcal{G}ain_k(M) = 1 - C_k(M)/C_k(M0)$$

Quantifying the change

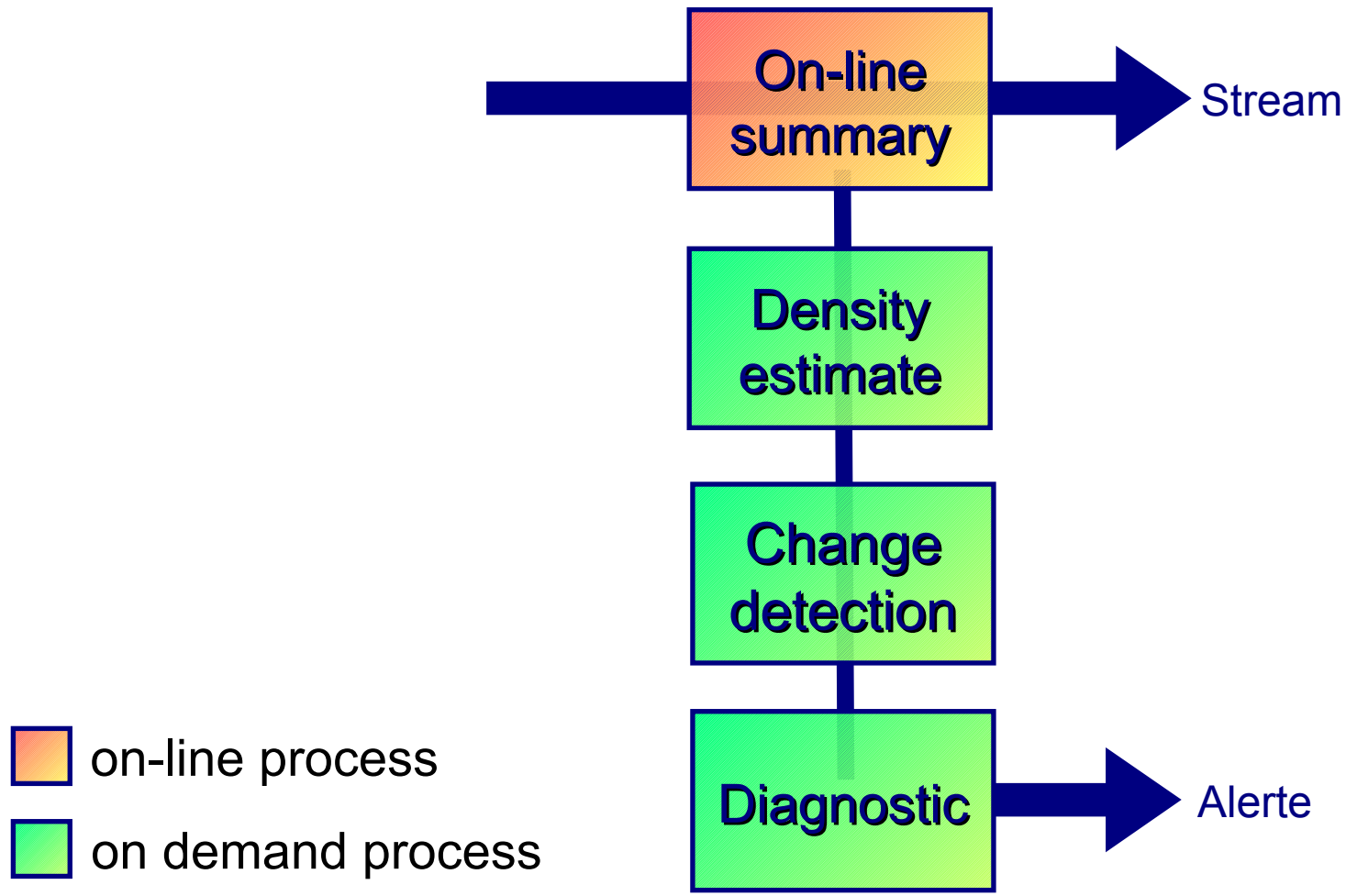
$$\mathcal{C}hange = \frac{1}{K} \sum_{k \in [1,K]} \mathcal{G}ain_k(\mathcal{M}ap)$$

Contribution indicator

$$\mathcal{C}ontrib(k) = \frac{\mathcal{G}ain_k(\mathcal{M}ap)}{K}$$

An alternative Approach

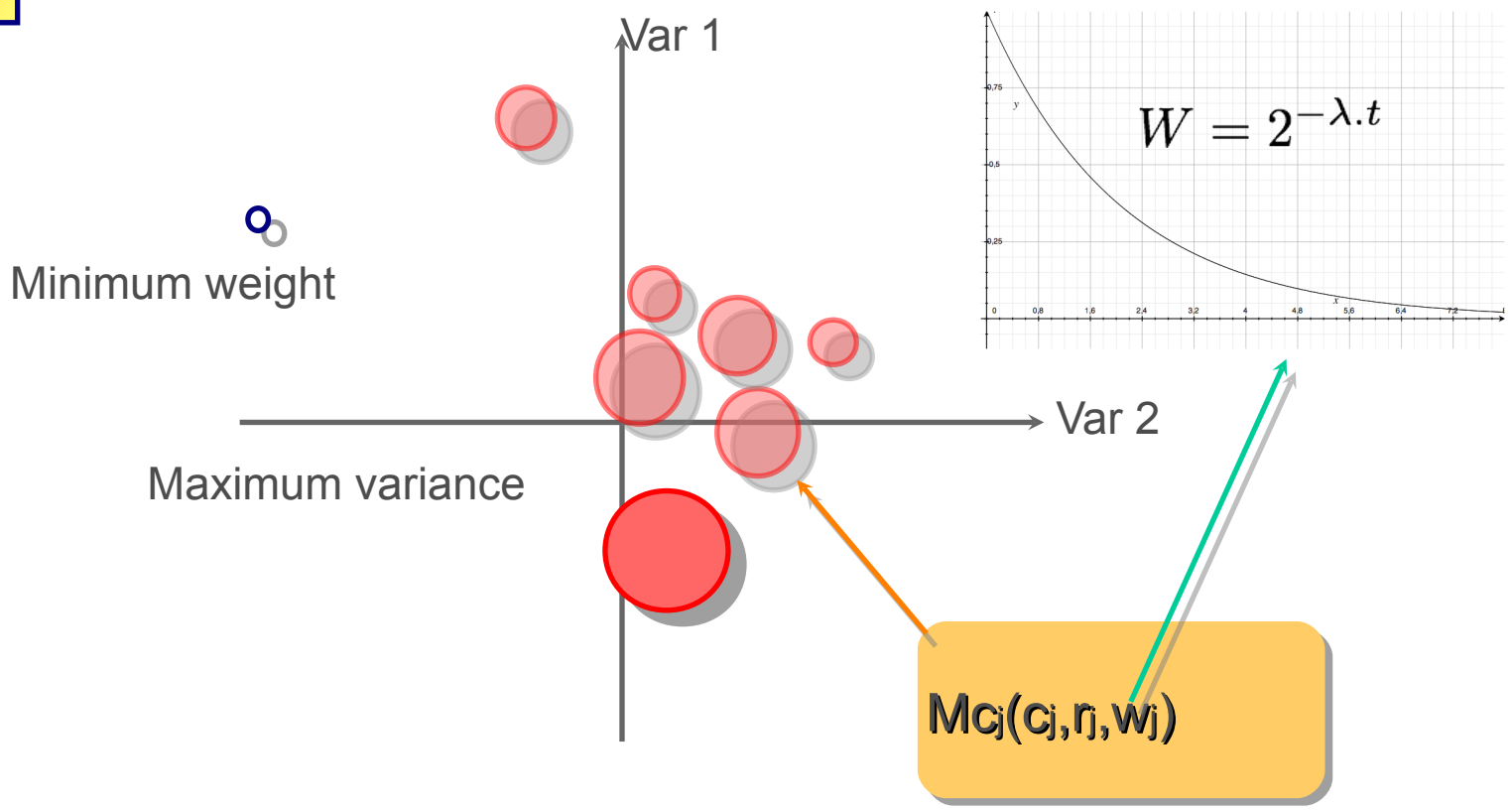
Four successive steps



An alternative Approach

On-line summary

Denstream Algorithm [Feng Cao 2006]



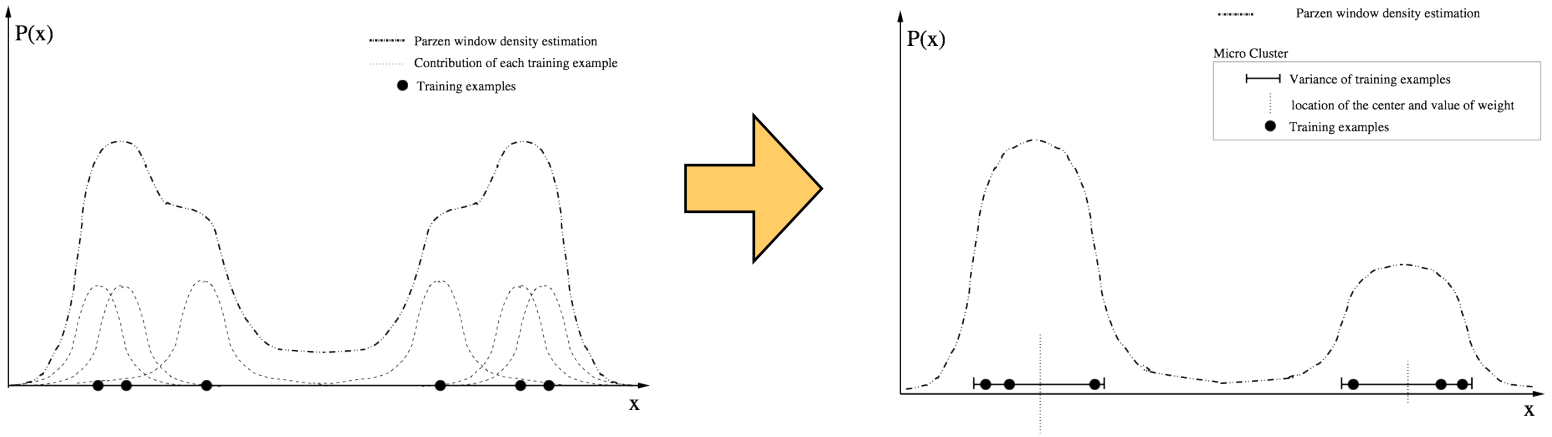
An alternative Approach

Density estimate

Adapted Parzen Window

$$\hat{P}^*(x) = \frac{1}{C.W} \sum_{j=1}^{C} \frac{\omega_j}{\sqrt{2\pi\left(\delta^2 + r_j^2\right)}^k} \exp^{-\frac{d(x,c_j)^2}{2\left(\delta^2 + r_j^2\right)}}$$

- W : total weight of the stream
- C : number of micro-clusters
- w_j : weight of the j^{th} micro-cluster
- r_j : standard deviation of the j^{th} micro-cluster
- delta : fading parameter



An alternative Approach

Change detection

Kullback-Leibler divergence

$$KL < P_{ref}(x) \| \hat{P}^*(x) >= - \int_{\mathbb{R}^k} P_{ref}(x) \log \frac{P_{ref}(x)}{\hat{P}^*(x)} dx$$

Diagnostic

The Kullback-Leibler divergence is evaluated within a sub-space *(excluding the ith variable)* KL^i_{minus}

$$Contrib(l) = \frac{\left(\sum_{i=1}^{k} KL^i_{minus}\right) - KL^l_{minus}}{\sum_{i=1}^{k} KL^i_{minus}}$$

An alternative Approach

Many parameters

On-line summary

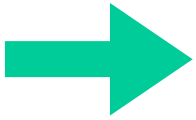
Denstream algorithm [Feng Cao 2006]

- fading parameter
- minimum weight micro-clusters
- maximum variance in micro-clusters

Density estimate

Modified version of Parzen windows

- smoothing parameter (gaussian kernel)



The tuning of these parameters is challenging in the case of data-streams

An alternative Approach

Curse of dimensionality

On-line summary

Denstream algorithm [Feng Cao 2006]

- micro-clusters are almost empty in case of high dimensionality

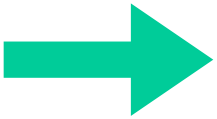
Change detection

Diagnostic

The exploited criteria is based on Kullback-Leibler divergence

$$KL < P_{ref}(x) \| \hat{P}^*(x) >= - \int_{\mathbb{R}^k} P_{ref}(x) \log \frac{P_{ref}(x)}{\hat{P}^*(x)} dx$$

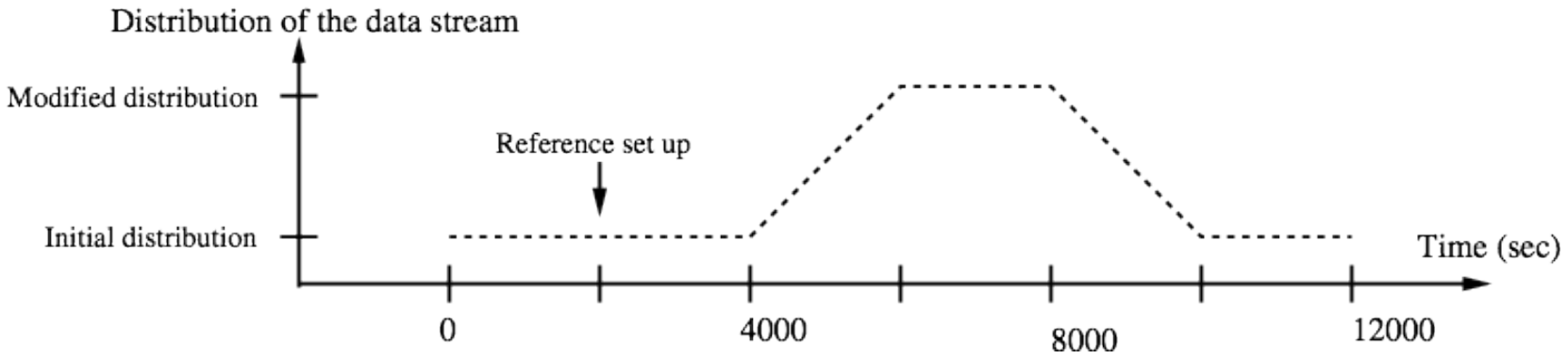
- The integral is estimated by the Monte-Carlo algorithm



time complexity and accuracy problems in case of high dimensionality

Comparative experiments

Artificial data-steams

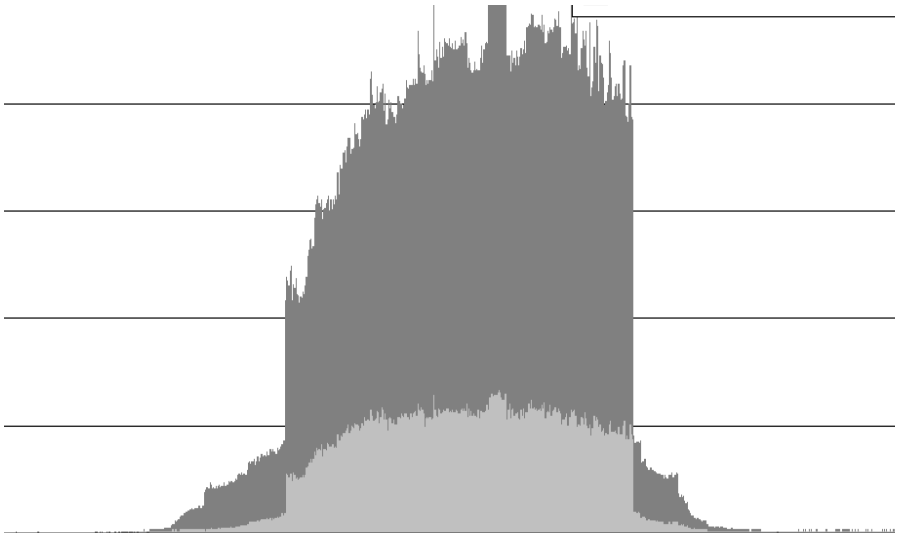


	initial distribution	modified distribution
Data stream 1 : change in mean	$\mathcal{N}\left(\begin{matrix}0 & 0\end{matrix}, \begin{matrix}1 & 0\\ 0 & 1\end{matrix}\right)$	$\mathcal{N}\left(\begin{matrix}4 & 8\end{matrix}, \begin{matrix}1 & 0\\ 0 & 1\end{matrix}\right)$
Data stream 2 : change in standard deviation	$\mathcal{N}\left(\begin{matrix}0 & 0\end{matrix}, \begin{matrix}1 & 0\\ 0 & 1\end{matrix}\right)$	$\mathcal{N}\left(\begin{matrix}0 & 0\end{matrix}, \begin{matrix}4 & 0\\ 0 & 9\end{matrix}\right)$

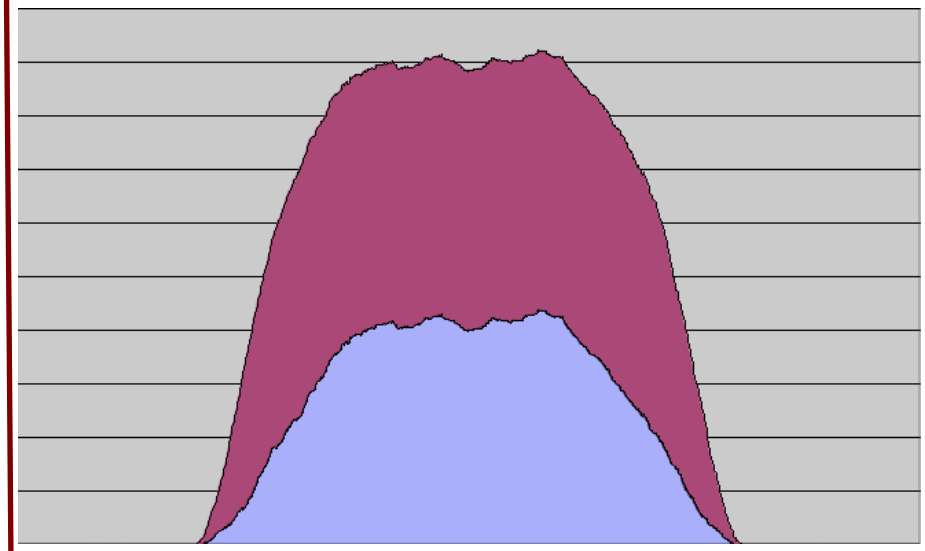
Comparative experiments

Change in mean

alternative method



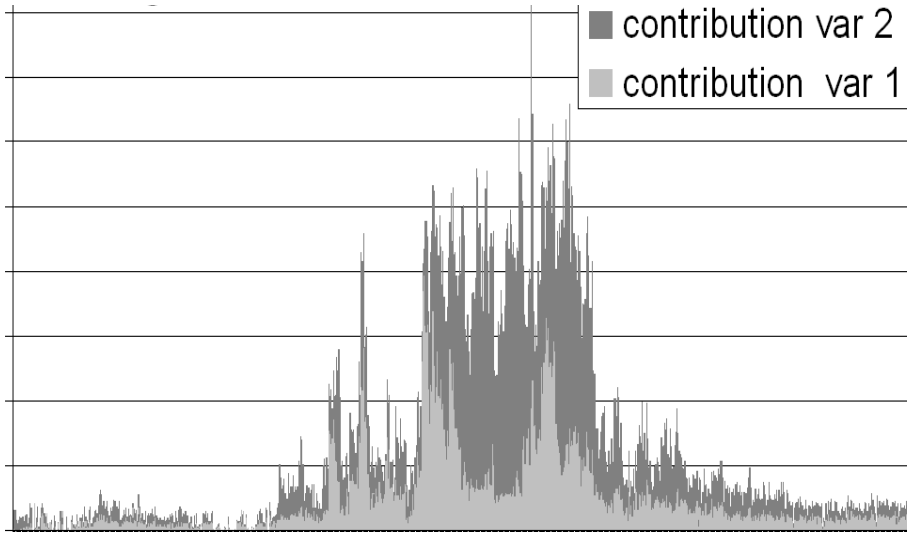
supervised method



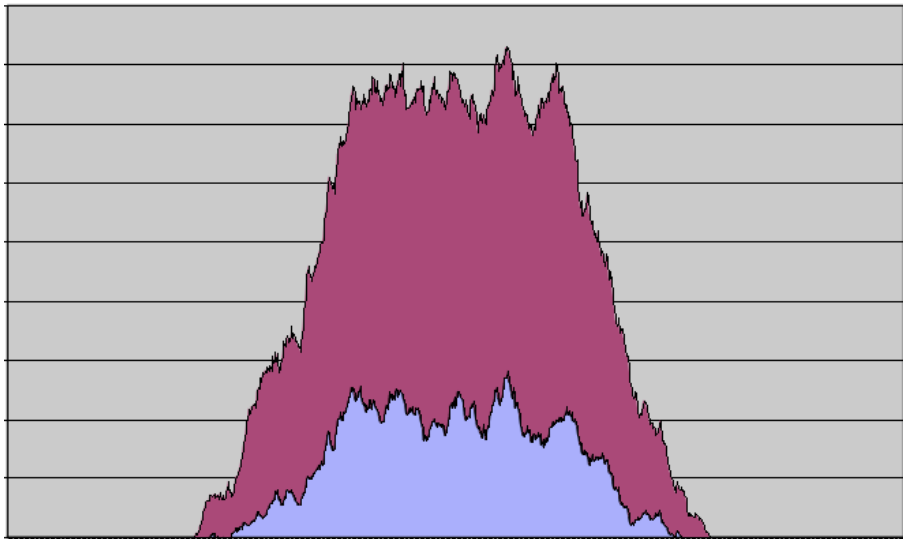
Comparative experiments

Change in standard deviation

alternative method



supervised method

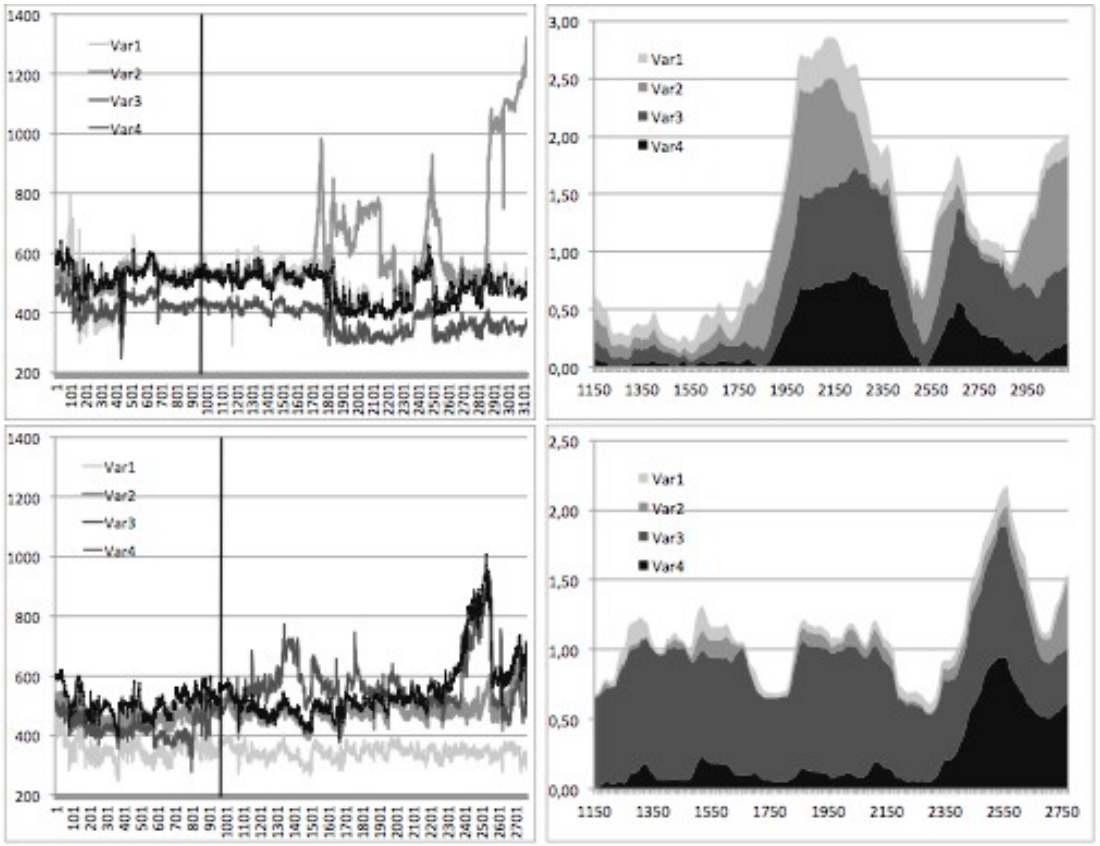


Conclusion (1/2)

- The supervised method provides the best results and is easier to implement :
  - No user parameters to be tuned
  - Only the choice of temporal windows

Conclusion (2/2)

- The supervised method gives promising results on real data streams:



Biblio

A. Bondu, M. Boullé, A. *« A Supervised Approach for Change Detection in Data Streams»*, IJCNN 2011

EGC 2013 tutorial: Data Grid Models

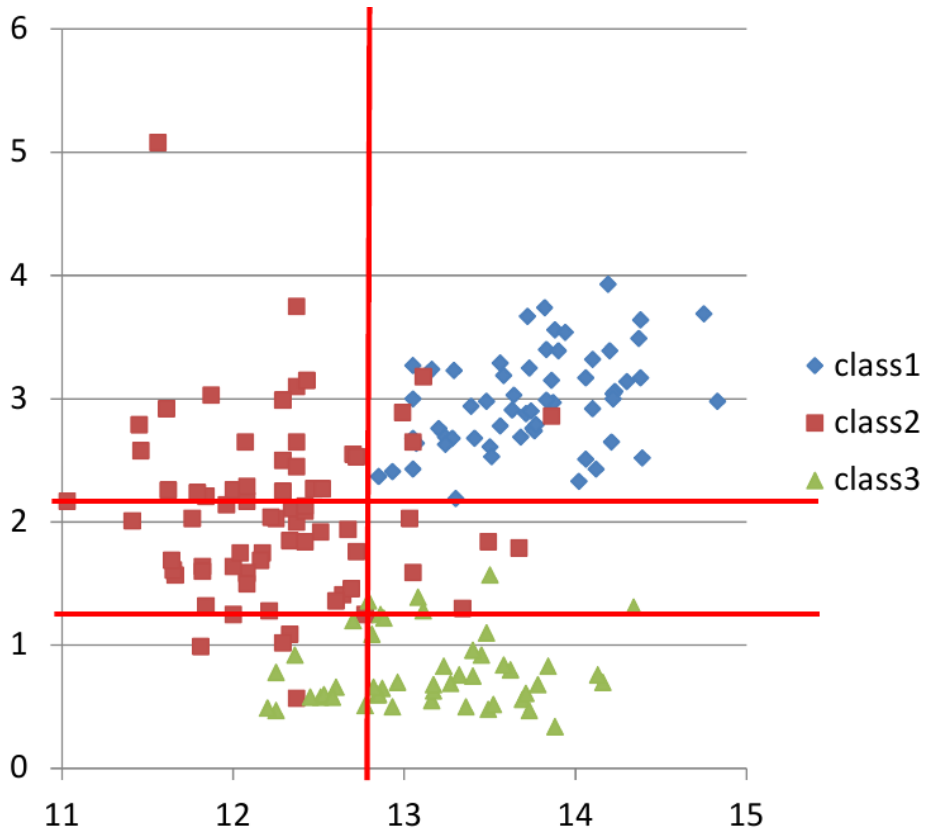
From data grid models to Classification Rules & decision Trees

Alexis Bondu, Marc Boullé, Dominique Gay

EDF R&D, Orange Labs

January 29, 2013

From data grid models. . .

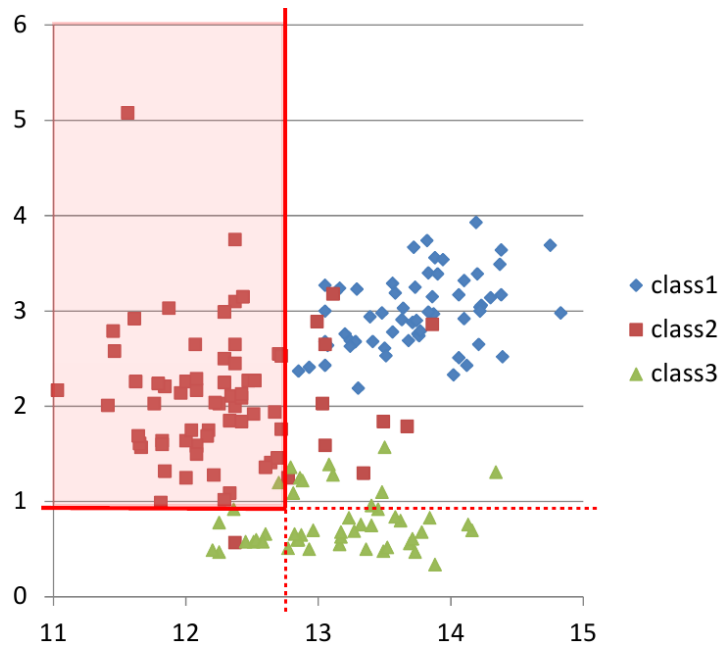


UCI-Wine data set: 3-class problem. 2-*d* data grid, 6 cells

. . . to Classification Rules & Decision Trees

Rules

Rule: $x \leq 12.8 \wedge y \geq 0.95 \rightarrow (0, 60, 1)$



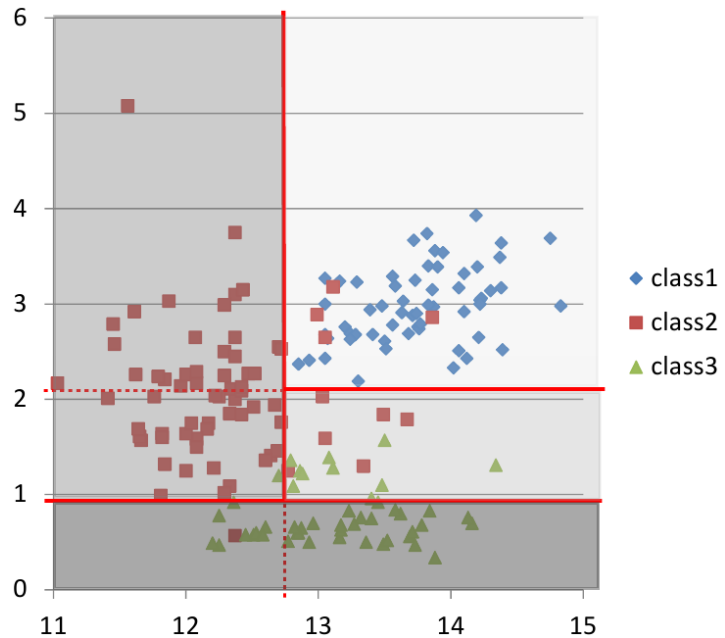
Tree

If $y \leq 0.95$ Then $(0, 1, 38)$

Else If $x \leq 12.8$ Then $(0, 60, 1)$

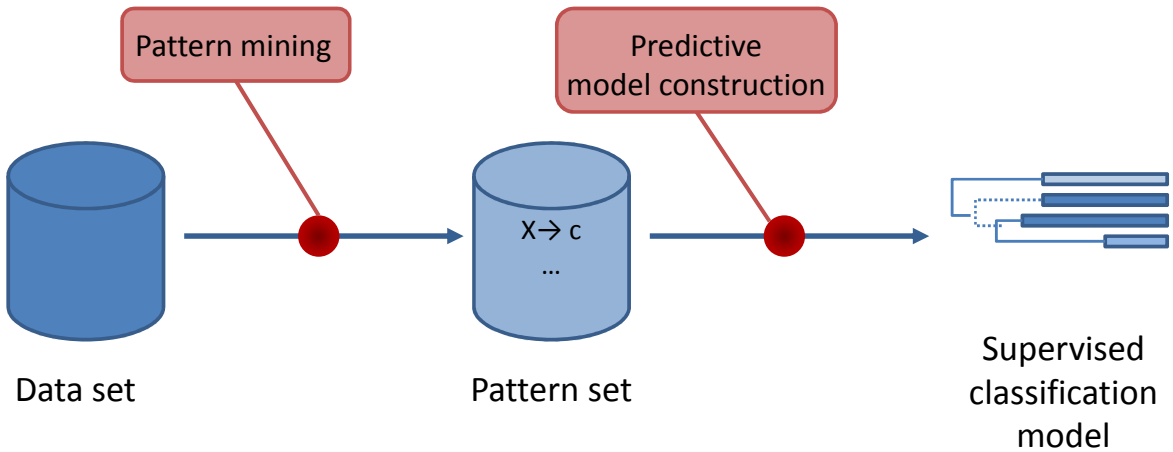
Else If $y \leq 2.1$ Then $(59, 4, 0)$

Else $(0, 6, 8)$



Context

Rule-based classification



survey [Bringmann et al. LeGo@ECML/PKDD'09]

Advantages

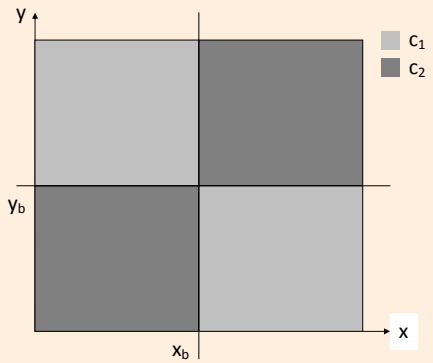
Good predictive performance and interpretability

Issues

State of the art

- Mostly dedicated to binary data
- Univariate preprocessing of non-binary attributes $\rightarrow$ Missing multivariate associations
- Many parameters to tune
- Non-robust pattern extraction
  ... without post-processing with parameters

XOR example



Parameter tuning

- Frequency threshold
- Interestingness measure threshold
- Number of rules (post-selection)
- Parameter for robust post-selection

Contributions

MODL rules (Minimum Optimized Description Length)

- Identify interesting and robust rules
- Handle binary/categorical/numeric attributes
- Find multivariate associations
- Parameter-free mining algorithm
- Anytime mining/classification process

Contents

Contents

The MODL approach

Data mining task $\simeq$ Model selection problem (Bayesian MAP approach)

Principle: Model selection betting on a trade-off between

- accuracy of the predictive information provided by the model
- robustness for a good generalization of the model

- model = discretization, value grouping, decision tree, data grid, . . . , rule
- The best model $M^\star$ (among the possible models $\mathcal{M}$) maximizes the a posteriori probability of a model M given data D :

$$M^\star = \arg\max_{M \in \mathcal{M}} p(M \mid D)$$

- MODL: Discrete model space
- MODL: Data-dependent hierarchical prior (uniform at each stage)

The best model is the most probable arising from the data

The MODL approach

Bayes theorem

$$p(M \mid D) = \frac{p(M) \times p(D \mid M)}{p(D)}$$

MODL evaluation criterion based on the cost of a model

$$c(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) = -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{likelihood}})$$

The best model is the model with least cost (shortest code length)

Processing MODL approach

- Define the model and the model space
- Establish the hierarchical prior on the model space
- Optimization algorithm to reach $M^\star$

MODL rules: the model space

Maximizing $p(\pi \mid D) = p(\pi) \times p(D \mid \pi)$

Minimizing $c(\pi) = -\log(p(\pi) \times p(D \mid \pi))$

Standard Classification Rule Model (SCRM)

MODL rule (SCRM) is uniquely defined by :

- the constituent attributes of the rule body
- the group involved in the rule body for each categorical body attribute
- the interval involved in the rule body for each numerical attribute
- the distribution of classes inside and outside of the body

Model space $\mathcal{M} = \{ \text{SCRM} \}$

Example

$\pi : (x_1 \in \{v_{x_1}^1, v_{x_1}^3, v_{x_1}^4\}) \wedge (1.2 \leq x_2 \leq 3.1) \wedge (x_4 \geq 100) \rightarrow (p_{c_1} = 0.9, p_{c_2} = 0.1)$

- group and interval items
- class distribution as consequent

MODL rules

Notations

Data set D with N objects, m attributes and J classes.
SCRM $\pi : X \rightarrow (p_{c_1}, p_{c_2}, \ldots, p_{c_J})$ such that $|X| = k \leq m$

- $X = \{x_1, \ldots, x_k\}$: the set of k constituent attributes of the rule body ($k \leq m$)
- $X_{cat} \cup X_{num} = X$: the sets of categorical and numerical attributes of the rule body
- $V_x = |dom(x)|$: the number of values of a categorical attribute x
- I_x: the number of intervals (resp. groups) of a numerical (resp. categorical) attribute x
- $\{i(v_x)\}_{v_x \in dom(x)}$: the indexes of groups to which v_x are affected (one index per value)
- $\{N_{i(x).}\}_{1 \leq i \leq I_x}$: the number of objects in interval i of numerical attribute x
- $i_{x_1}, \ldots, i_{x_k}$: the indexes of groups of categorical attributes (or intervals of numerical attributes) involved in the rule body

Contingency table

π	c_1	$\ldots$	c_J	$\sum$
X	N_{X_1}	$\ldots$	N_{X_J}	N_X
$\neg X$	$N_{\neg X_1}$	$\ldots$	$N_{\neg X_J}$	$N_{\neg X}$
$\sum$	N_1	$\ldots$	N_J	N

MODL rules: prior distribution of the model space

Hierarchical prior (data-dependent and uniform at each stage)

- *(i)* the number of attributes in the rule body is uniformly distributed between 0 and m
- *(ii)* for a given number k of attributes, every set of k constituent attributes of the rule body is equiprobable
- *(iii)* for a given categorical attribute in the body, the number of groups is necessarily 2
- *(iv)* for a given categorical attribute, for a value group of this attribute, belonging to the body or not are equiprobable
- *(v)* for a given numerical attribute in the body, the number of intervals is either 2 or 3 (with equiprobability)
- *(vi)* for a given numerical attribute with 2 intervals, for an interval of this attribute, belonging to the body or not are equiprobable. When there are 3 intervals, the body interval is necessarily the middle one
- *(vii)* for a given categorical (or numerical) attribute, for a given number of groups (or intervals), every partition of the attribute into groups (or intervals) is equiprobable
- *(viii)* every distribution of the class values is equiprobable, in and outside of the body
- *(ix)* the distributions of class values in and outside of the body are independent

MODL rules: probabilities

$$p(\pi \mid D) = p(\pi) \times p(D \mid \pi)$$

Prior: $p(\pi)$

$$p(\pi) = p(X) \qquad \text{// constituent attributes}$$

$$\times \prod_{X_{cat}} p(I_x) \times p(\{i(v_x)\}|I_x) \times p(i_x|\{i(v_x)\}, I_x) \qquad \text{// categorical attributes}$$

$$\times \prod_{X_{num}} p(I_x) \times p(\{N_{i(x).}\}|I_x) \times p(i_x|\{N_{i(x).}\}, I_x) \qquad \text{// numerical attributes}$$

$$\times \prod_{i \in \{X, \neg X\}} p(\{N_{ij}\} \mid N_X, N_{\neg X}) \qquad \text{// class distribution in/out body}$$

MODL rules: probabilities

$$X_{cat}: \quad p(I_x) \times p(\{i(v_x)\}|I_x) \times p(i_x|\{i(v_x)\}, I_x)$$

$$1 \times \frac{1}{\mathcal{S}(V_x, 2)} \times \frac{1}{2}$$

$$X_{num}: \quad p(I_x) \times p(\{N_{i(x).}\}|I_x) \times p(i_x|\{N_{i(x).}\}, I_x)$$

$$\frac{1}{2} \times \frac{1}{\binom{N-1}{I_x-1}} \times \frac{1}{1 + \mathbb{1}_{\{2\}}(I_x)}$$

$$\prod_{i \in \{X, \neg X\}} p(\{N_{ij}\} \mid N_X, N_{\neg X})$$

$$\frac{1}{\binom{N_X+J-1}{J-1}} \times \frac{1}{\binom{N_{\neg X}+J-1}{J-1}}$$

Likelihood:

$$p(D|\pi)$$

$$\frac{1}{\frac{N_X!}{\prod_{j=1}^{j=J} N_{Xj}!}} \times \frac{1}{\frac{N_{\neg X}!}{\prod_{j=1}^{j=J} N_{\neg Xj}!}}$$

MODL rules: interestingness measure

Cost of a rule: $c(\pi) = -\log(p(\pi) \times p(D \mid \pi))$

$$c(\pi) = \log(m+1) + \log \binom{m+k-1}{k} \tag{1}$$

$$+ \sum_{X_{cat}} \log \mathcal{S}(V_x, 2) + \log 2 + \sum_{X_{num}} \log 2 + \log \binom{N-1}{I_x - 1} + \log(1 + \mathbb{1}_{\{2\}}(I_x)) \tag{2}$$

$$+ \log \binom{N_X + J - 1}{J-1} + \log \binom{N_{\neg X} + J - 1}{J-1} \tag{3}$$

$$+ \left(\log N_X! - \sum_{j=1}^{j=J} \log N_{Xj}! \right) + \left(\log N_{\neg X}! - \sum_{j=1}^{j=J} \log N_{\neg Xj}! \right) \tag{4}$$

level: normalized evaluation criterion

$$level(\pi) = 1 - \frac{c(\pi)}{c(\pi_\emptyset)}$$

$$c(\pi_\emptyset) = \log(m+1) + \log \binom{N+J-1}{J-1} + \log N! - \sum_{j=1}^{j=J} \log N_j!$$

MODL rules: behavior

Interpretation: $level(\pi) = 1 - c(\pi)/c(\pi_\emptyset)$

- $level(\pi) \leq 0$: less probable than the default rule
- $level(\pi) > 0$: probable rule bringing predictive information

$c(\pi_\emptyset) \rightarrow N \times H(y)$

$$\lim_{N\to\infty} \frac{c(\pi_\emptyset)}{N} = -\sum_{j=1}^{j=J} \frac{N_j}{N} \log \frac{N_j}{N}$$

$c(\pi) \rightarrow N \times H(y|X)$

$$\lim_{N\to\infty} \frac{c(\pi)}{N} = \frac{N_X}{N}\left(\sum_{j=1}^{j=J} -\frac{N_{Xj}}{N_X} \log \frac{N_{Xj}}{N_X}\right) + \frac{N_{\neg X}}{N}\left(\sum_{j=1}^{j=J} -\frac{N_{\neg Xj}}{N_{\neg X}} \log \frac{N_{\neg Xj}}{N_{\neg X}}\right)$$

Interpretation

- *level*: class entropy ratio
- $level(\pi) \leq 0$: not significant patterns (arising from randomness)

MODL rules : problem formulation

Size of the model space

$$O((2^{V_c})^{m_c}(N^2)^{m_n})$$

- m_c number of categorical attributes with V_c values
- m_n the number of numerical attributes

No exhaustive mining

Simpler formulation

Efficiently mining with diversity a set of SCRM with *level* ≥ 0

Contents

Mining algorithm

Principle

Randomized strategy for sampling the posterior distribution of SCRM rules

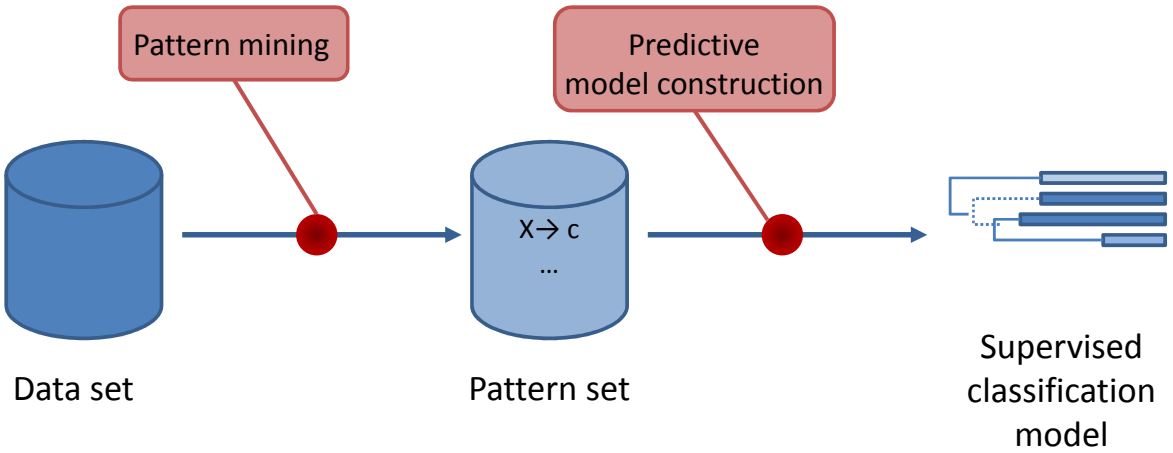
Main algorithm: MACATIA

1: **repeat**
2: $t \leftarrow$ *chooseRandomObject*($\mathcal{T}$)
3: $I \leftarrow$ *chooseRandomAttributes*($\mathcal{I}$)
4: $X \leftarrow$ *chooseRandomCoveringItemSet*(t, I)
5: $\pi \leftarrow$ *optimizeRule*(t, I)
{Moving intervals bounds}
{Changing value groups}
6: **until** *timeStoppingCondition*

Complexity : Mining one rule in $O(kN \log N)$

- randomized
- instance-based
- anytime
- parameter-free
- locally optimal

Classification system



KRSNB: principle

Simple feature construction process (ended with Selective Naive Bayes SNB (Boullé JMLR'07))

New feature space

For each mined rule π, a new Boolean feature f is built,

- $t(f) = 1$ if t supports π body
- $t(f) = 0$ otherwise

Contents

Protocol & data sets

UCI benchmark data

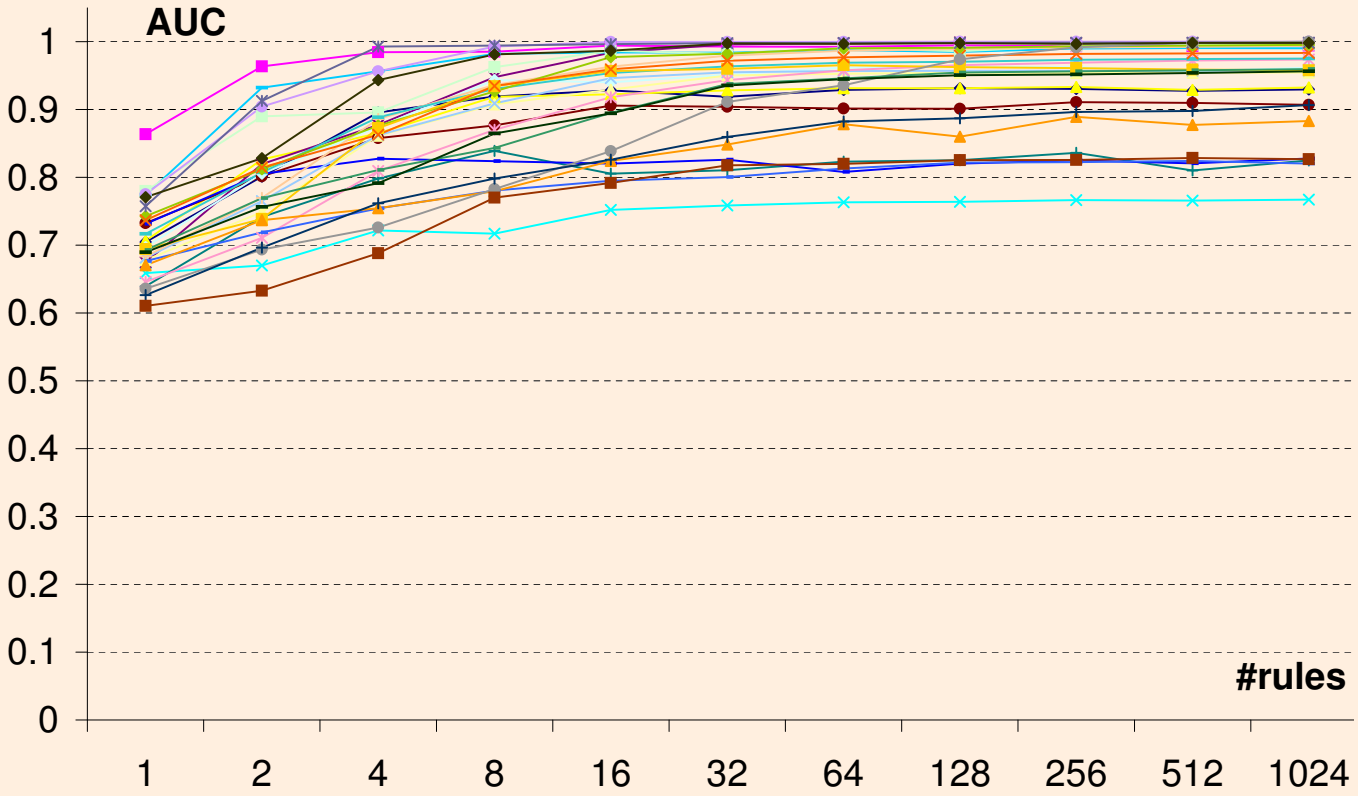
- from 150 to 20000 instances
- from 4 to 60 attributes (various types)
- from 2 to 26 classes (some imbalanced)
- 10-folds cross validation

Real-world challenge data set

- Orange KDD 2009
  - 50000 instances, 230 (190 numerical, 40 categorical) variables
  - 2 classes (highly imbalanced, 98/02 or 92/08 depending on task)
- Neurotech PAKDD 2009 & 2010
  - 50000 instances, 31-53 variables
  - 2 classes (imbalanced, 2009: 80/20 ; 2010: 76/24)
- 70% train 30% test experiments

Experimental Validation (efficiency)

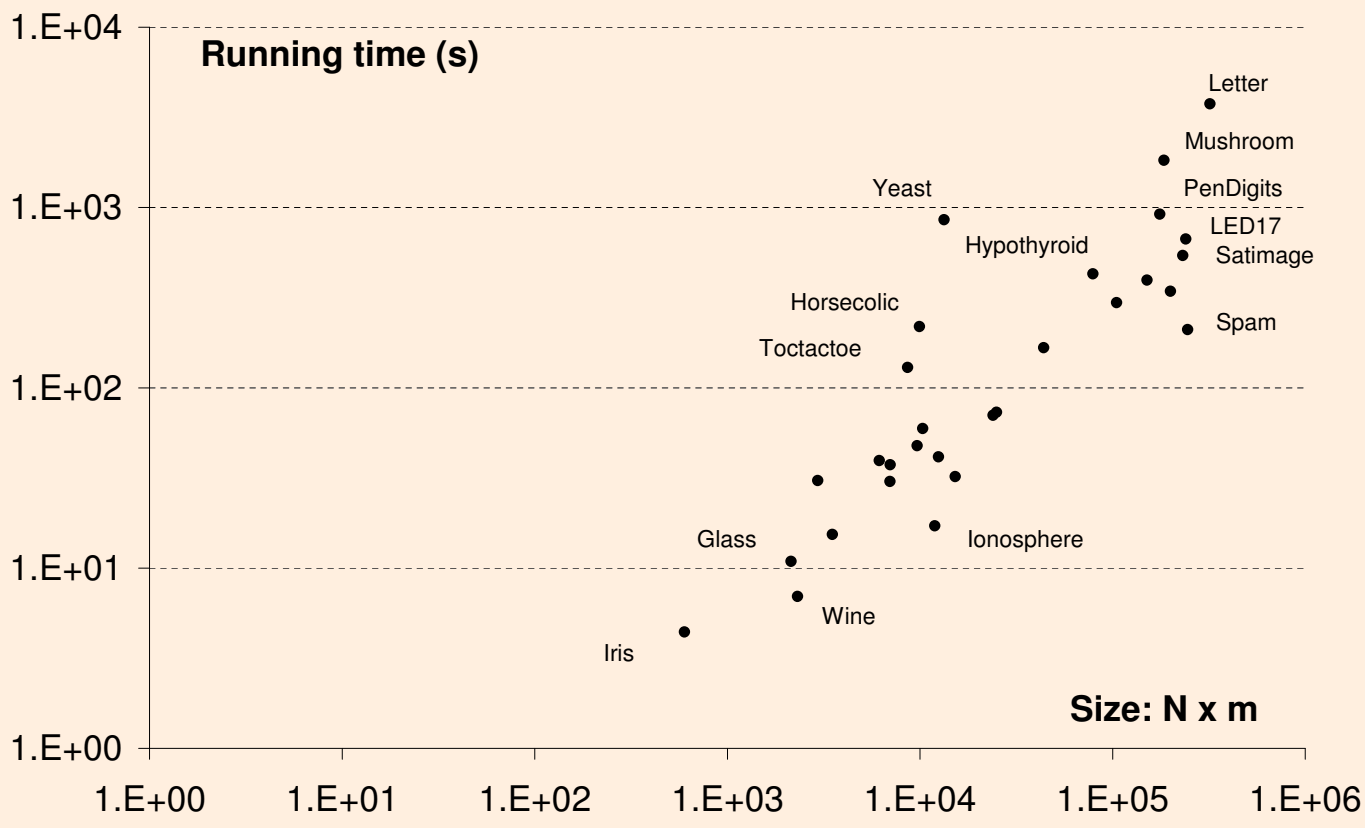
Performance evolution



Adding a few rules as new features increases predictive performance

Experimental Validation (time efficiency)

Time efficiency (for mining 1024 rules)



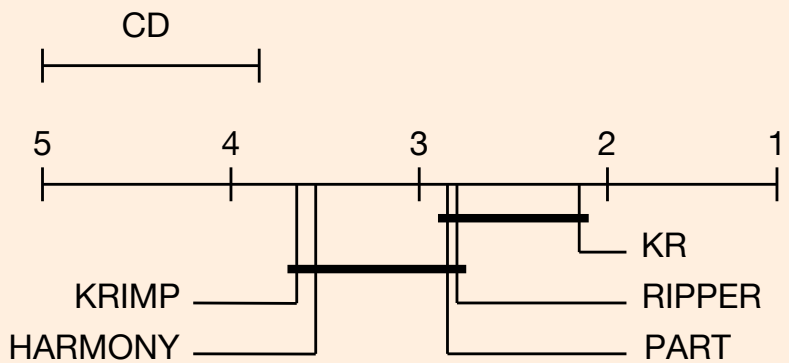
Reaching top performance with few rules in reasonable time

Predictive performance (competitivity)

KRSNB versus state-of-the-art

Algorithms	avg.acc	avg.rank	KR-WTL
KRSNB	**84.80**	**2.17**	-
HARMONY	83.31	3.53	**19**/1/9
KRIMP	83.31	3.64	**23**/1/5
RIPPER	84.38	2.83	**19**/1/9
PART	84.19	2.83	**18**/1/10

Critical difference diagram



KRSNB $>$ KRIMP,HARMONY ; KRIMP,HARMONY $\approx$ RIPPER,PART

KRSNB highly competitive

Large-scale challenge data set

Pre-processing by discretization/binarization makes the task unfeasible unless pruning numerous attributes

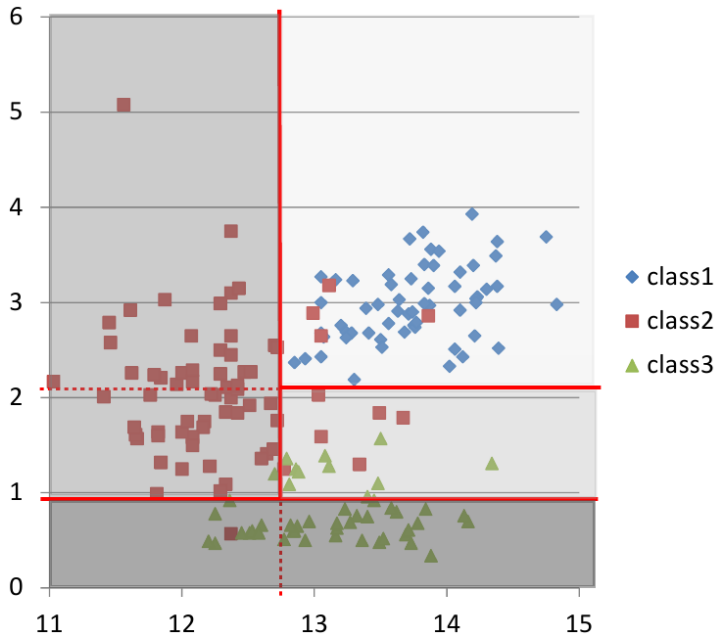
AUC results

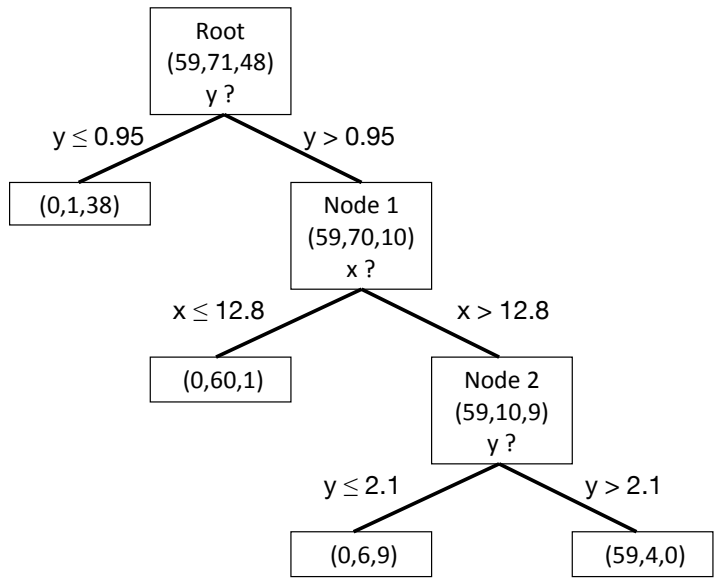
	NEUROTECH-PAKDD		ORANGE-KDD'09		
	2009	2010	APPET.	CHURN	UPSELL.
KRSNB	**66.31**	**62.27**	**82.02**	**70.59**	**86.46**
RIPPER	51.90	50.70	50.00	50.00	71.80
PART	59.40	59.20	76.40	64.70	83.50

KRSNB is highly competitive on real large-scale data set

Contents

Decision tree example





Tree:

$$\text{If } y \leq 0.95 \text{ Then } (0, 1, 38)$$
$$\text{Else If } x \leq 12.8 \text{ Then } (0, 60, 1)$$
$$\text{Else If } y \leq 2.1 \text{ Then } (59, 4, 0)$$
$$\text{Else } (0, 6, 8)$$

MODL trees: the model space

Maximizing $p(\tau \mid D) = p(\tau) \times p(D \mid \tau)$

Minimizing $c(\tau) = -\log(p(\tau) \times p(D \mid \tau))$

MODL tree

MODL tree is uniquely defined by :

- its structure
  - the constituent attributes of the tree
  - the nature of nodes (internal or leaves)
- the repartition of the objects in this structure
  - the groups/intervals of attributes in internal nodes
  - the repartition of objects in internal nodes
  - the class distribution in the leaves

MODL tree criterion

Cost of a tree: $c(\tau) = -\log(p(\tau) \times p(D \mid \tau))$

$$c(\tau) = \log(m+1) + \log \binom{m+k-1}{k} \tag{5}$$

$$+ \sum_{s \in \mathcal{S}_{T_n}} \log k + C_{Ris}(I_s) \log 2 + \log \binom{N_{s.} + I_s - 1}{I_s - 1} \tag{6}$$

$$+ \sum_{s \in \mathcal{S}_{T_c}} \log k + C_{Ris}(I_s) \log 2 + \log B(V_{X_s}, I_s) \tag{7}$$

$$+ \sum_{l \in \mathcal{L}_T} C_{Ris}(1) \log 2 + \log \binom{N_{l.} + J - 1}{J - 1} \tag{8}$$

$$+ \sum_{l \in \mathcal{L}_T} \log \frac{N_{l.}!}{N_{l.1}! N_{l.2}! \ldots N_{l.J}!} \tag{9}$$

Learning algorithm

Principle

Classical Top-down construction of the tree
Two strategies:

- pre-pruning (MT)
  - growing tree while improving global criterion
  - choosing the best attribute and MODL 1-D on nodes
- post-pruning (MTp)
  - growing tree while there MODL informative variables
  - post-pruning nodes if improving global criterion
- binary (2) vs n-ary trees

Complexity : $O(mJN^2 \log N)$

- deterministic
- parameter-free
- locally optimal

Experiments

Experiments on UCI data and WCCI challenge data

- The better the criterion, the more predictive the tree
- Binary trees are better
- Predictive performance: KT $\simeq$ C4.5, CART
- Complexity/Size of tree : KT produces simpler trees

Method	Train data set Acc.	Train data set AUC	Test data set Acc.	Test data set AUC	Size	Time	$C_{opt}(T)$
MT(2)	0.845	0.914	0.819	0.889	17.5	0.5	524
MT	0.841	0.915	0.813	0.884	19.4	0.5	565
MTp(2)	0.840	0.910	0.822	0.891	17.4	0.6	508
MTp	0.834	0.905	0.817	0.890	19.5	0.6	547
NMT	0.879	0.959	0.762	0.844	142.3	0.8	1095
sCART	0.854	0.921	0.822	0.876	30.7	1.0	×
J48	0.929	0.962	0.834	0.881	77.1	0.1	×

Data Set	Accuracy MTp(2)	Accuracy MTp	Accuracy sCart	Accuracy J48	AUC MTp(2)	AUC MTp	AUC sCart	AUC J48	Tree Size MTp(2)	Tree Size MTp	Tree Size sCart	Tree Size J48
ada	0.847	0.847	0.842	0.846	0.887	0.890	0.860	0.860	22.0	23.6	28.1	224.0
gina	0.881	0.863	0.894	0.867	0.923	0.913	0.918	0.862	47.8	49.1	64.4	247.7
hiva	0.966	0.966	-	0.955	0.622	0.622	-	0.659	6.0	6.0	-	64.4
nova	0.866	0.866	-	-	0.817	0.817	-	-	17.6	17.6	-	-
sylva	0.989	0.989	0.991	0.990	0.991	0.991	0.981	0.954	26.2	41.4	41.0	105.2

Relevance of the criterion and Good predictive performance with simple trees

Contents

Conclusion

Summary

Mining classification rules

- in quantitative large-scale data sets
- identify interesting and robust rules
- parameter-free,competitive mining/classification process

Building decision trees

- parameter-free
- simple trees
- competitive predictive performance

Perspectives

- extension for regression rules, descriptive association rules
- extension to regression trees and forest

Data grid models
Conclusion

Alexis Bondu, Marc Boullé, Dominique Gay

January, 29, 2013

Schedule

- 14h15 : Data grid models
  - Principles, evaluation, optimisation
- 15h15 : Data Grid Models for Coclustering
  - Focus on model selection
- 16h15 : Pause (30 min)
- 16h45 : Coclustering applications using data grid models
  - Clustering of text, graph, text, curves, web logs…
- 17h15 : Data grid models for supervised learning
  - Application to data preparation and to change detection in stream mining
- 17h45 : Extension of data grid models
  - Classification rules and decision trees
- 18h30 : Conclusion
  - Summary, future work, discussion

MODL approach
Summary

- Data grid models for non parametric density estimation
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Data grid based on the cross-product of the univariate partitions, with a piecewise constant density estimation in each cell of the grid
  - Bayesian approach for model selection
  - Efficient optimization algorithms

- Model selection approach
  - Similar to Bayesian or MDL model selection
  - Model of the finite data sample
  - Asymptotical convergence to the true distribution when it exist
    - Proof in the case of coclustering of two categorical variables
    - Open question in the other cases

MODL approach
Extension and future work

- Generalization of the MODL approach
  - Partition the input representation
  - Partition the output representation
  - In each input part, describe the distribution of the output parts
- Application to alternative modeling techniques
  - K-nearest neighbours
  - Decision trees
  - Decision rules
- Application to alternative representations (other than data table)
  - Distance matrix
  - Graph
  - Time series
  - Relational database
  - Feature construction
- Theoretical foundations
  - Data dependent model space and prior
  - Proof of asymptotic consistency in the categorical case
  - Open questions
    - Asymptotic consistency in the general case
    - Convergence rate

MODL approach
New in EGC 2013

- Feature construction — January, 30, 2013, session 1.2, 11h00
  - Vers une Automatisation de la Construction de Variables pour la Classification Supervisée, M. Boullé, D. Lahbib
- Multi-table relational data mining — January, 30, 2013, session 1.2, 11h00
  - Un Critère d'Évaluation pour la Construction de Variables à base d'Itemsets pour l'Apprentissage Supervisé Multi-Tables, D. Lahbib, M. Boullé, D. Laurent
- Segmentation of call detail records — February, 1, 2013, session 6.1, 10h30
  - Étude des corrélations spatio-temporelles des appels mobiles en France, R. Guigourès, M. Boullé, F. Rossi
- Change detection in supervised stream mining — February, 1, 2013, session 6.2, 10h30
  - Grille bivariée pour la détection de changement dans un flux étiqueté, C. Salperwyck, M. Boullé, V. Lemaire
- Supervised classification of time series — February, 1, 2013, session 6.2, 10h30
  - Construction de descripteurs à partir du coclustering pour la classification supervisée de séries temporelles , D. Gay, M. Boullé
- Clustering of paths in a network — February, 1, 2013, session 6.2, 10h30
  - Classifications croisées de données de trajectoires contraintes par un réseau routier, M. K. El Mahrsi, R. Guigourès, F. Rossi, M. Boullé

MODL approach
Contact

- Tool available as a shareware
  http://www.khiops.com

- Contact
  - **Alexis Bondu**
    - EDF R&D
    - alexis.bondu@edf.fr
    - http://alexisbondu.free.fr/
  - **Marc Boullé**
    - Orange labs
    - marc.boulle@orange.com
    - http://perso.rd.francetelecom.fr/boulle/
  - **Dominique Gay**
    - Orange labs
    - dominique.gay@orange.com
    - https://sites.google.com/site/dominiquehomepage/home

Thank you for your attention!