Empirical comparison of a monothetic divisive clustering method with the Ward and the k-means clustering methods

Marie Chavent, Yves Lechevallier

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,

Université Bordeaux1, 351, Cours de la libération, 33405 Talence Cedex, France

chavent@math.u-bordeaux1.fr

Institut National de Recherche en Informatique et en Automatique,

Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France

Yves.Lechevallier@inria.fr

PART 1

INTRODUCTION

Introduction

- DIVCLUS-T is a descendant hierarchical clustering method
- it is designed for numerical or categorical data
- like WARD and the k-means, it is based on the minimization of the inertia criterion
- unlike WARD and the k-means, it provides by construction a simple and natural interpretation of the clusters

QUESTION: What is the price paied, in term of inertia, for this additional interpretation ?

Introduction

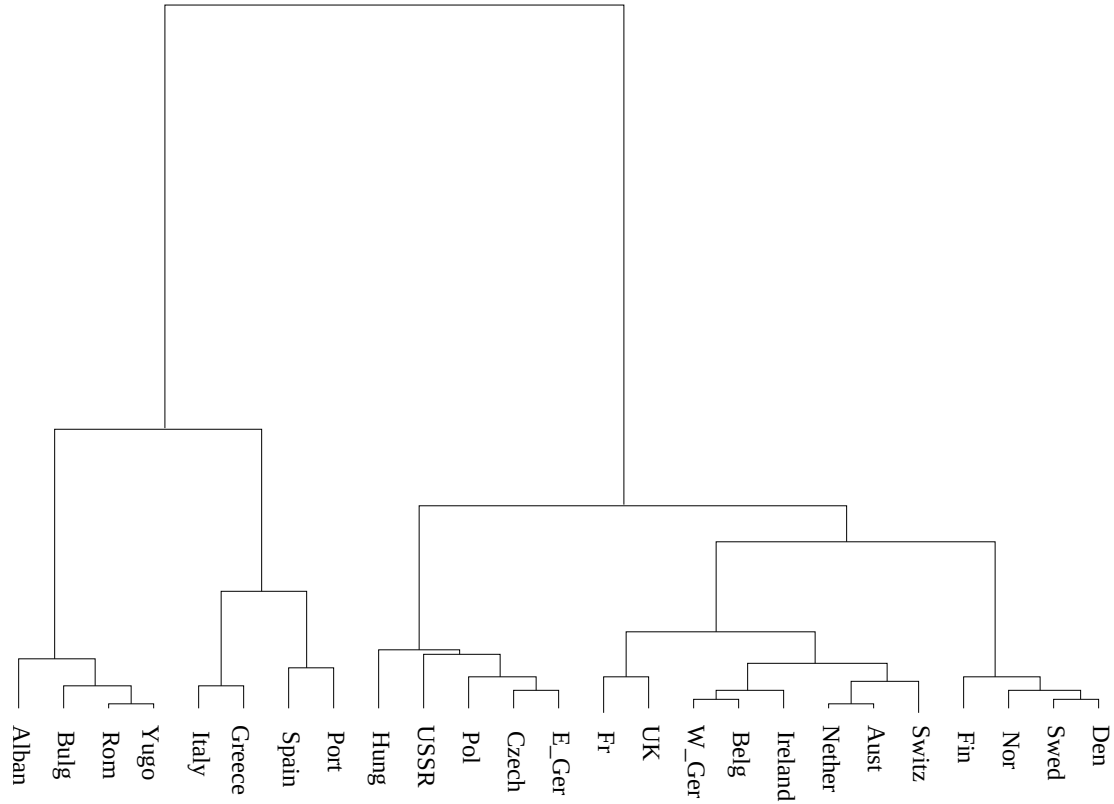
A numerical example : the Protein data

Country	Red Meat	White Meat	Eggs	Milk	Fish	Starchy Cereals	Foods	Nuts	Fruit/Veg.
Alban	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Aust	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belg	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulg	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czech	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Den	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
Finl	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
E-Ger	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Fr	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hung	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Nether	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Nor	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Pol	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Port	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Rom	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Swed	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switz	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W-Ger	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugo	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

A Handbook of Small Data Sets, Hand, D.J. et al. (eds.) (1994)

Introduction

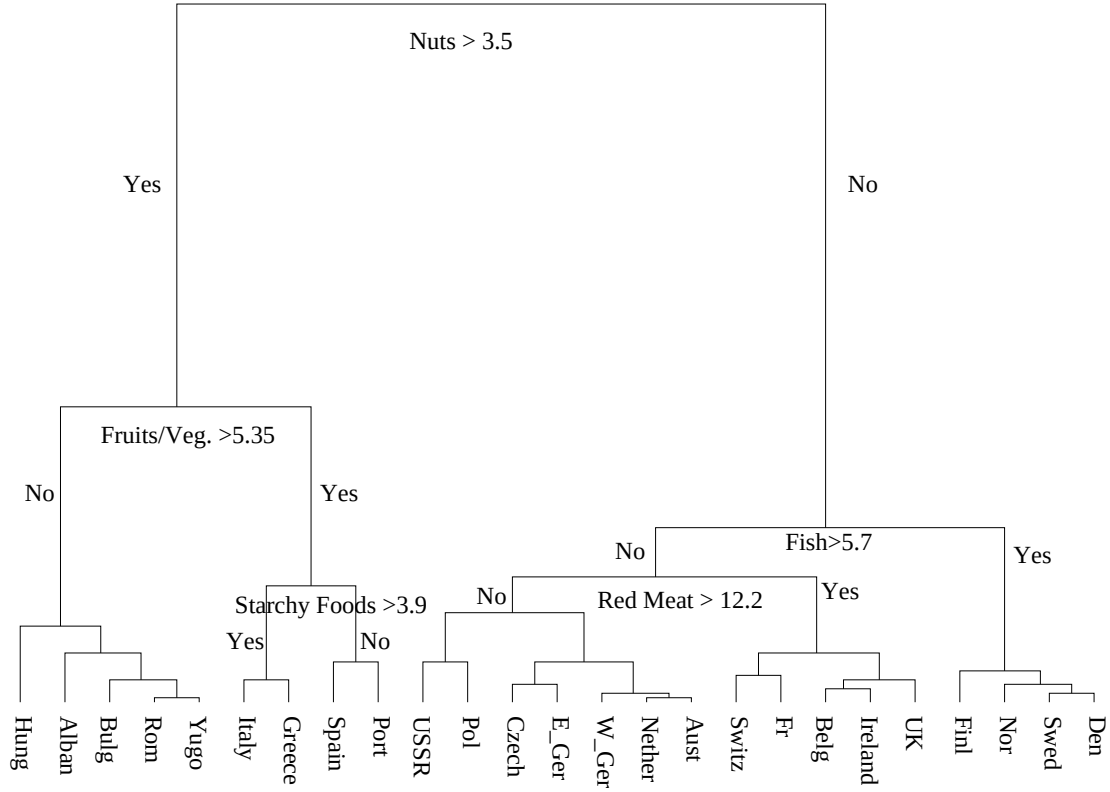
WARD dendrogram:



- the dendrogram gives no informations about the interpretations of the clusters
- example : {Fin, Nor, Swed, Den}

Introduction

DIVCLUS-T dendrogram:



Advantage: the clusters have natural interpretations

Question: How do the within-cluster inertia of the partitions obtained with WARD and DIVCLUS-T compare ?

Introduction

Percentages of inertia explained by the partitions of the WARD and the DIVCLUS-T dendrograms :

k	2	3	4	5	6	7	8	9	10
DIVCLUS-T	37.1	50.6	59.2	65.5	71.2	73.5	79.3	81.6	84
WARD	34.7	48.5	58.5	66.7	72.4	75.5	79	81.6	84

PART 2

The inertia criterion

The inertia criterion

X numerical

- The inertia of a cluster C_l :

$$I(C_l) = \sum_{i \in C_l} w_i \parallel x_i - g_l \parallel^2_M$$

$\Rightarrow$ measures the heterogeneity of a cluster

$\Rightarrow$ $M = I$ ou $M = D_{1/\sigma^2}$.

$\Rightarrow$ $w_i = 1$: SSQ criterion

- The within-cluster inertia of a partition $P_k = (C_1, ..., C_k)$

$$W(P_k) = \sum_{l=1}^{k} I(C_l)$$

$\Rightarrow$ measures the heterogeneity of a partition

The inertia criterion

$X_{n\times p}$ categorical, $w_i = 1$

Categorical example $X_{27\times 7}$: “Dogs” data (Saporta 1990)

Race	Size	Weight	Swiftness	intelligence	Affection	aggressiveness	Fonction	w_i
Beauceron	large	heavy	fast	fairly intel	very affect	aggress	utility	1
Basset Hound	small	light	slow	not very intel	not very affect	aggress	hunting	1
German Shepherd	large	heavy	very fast	very intel	very affect	aggress	utility	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Levrier	large	heavy	very fast	not very intel	not very affect	nonaggress	hunting	1
Mastiff	large	very heavy	slow	not very intel	not very affect	aggress	utility	1
Pekingese	small	light	slow	not very intel	very affect	nonaggress	company	1
Pointer	large	heavy	very fast	very intel	not very affect	nonaggress	hunting	1
Saint-Bernard	large	very heavy	slow	fairly intel	not very affect	aggress	utility	1
Setter	large	heavy	very fast	fairly intel	not very affect	nonaggress	hunting	1
Teckel	small	light	slow	fairly intel	very affect	nonaggress	company	1
Newfoundland	large	very heavy	slow	fairly intel	not very affect	nonaggress	utility	1
# of categories	2	3	3	2	2	2	3	

$\tilde{X}_{n\times q}$ numerical, $q = \sum_{j=1}^{p} m^j$, $w_i = 1/n$

Categorical example $\tilde{x}_{27\times 17}$

The inertia criterion

Transforming $X_{n\times p}$ (categorical) in $\tilde{X}_{n\times q}$ (numerical)

- $X_{n\times p} \rightarrow$ indicator matrix $\mathcal{K}_{n\times q}$ with 0 or 1 entries

The inertia criterion

Transforming $X_{n\times p}$ (categorical) in $\tilde{X}_{n\times q}$ (numerical)

- $X_{n\times p} \rightarrow$ indicator matrix $\mathcal{K}_{n\times q}$ with 0 or 1 entries
- $\mathcal{K}_{n\times q}$ contingency table $\rightarrow$ row profiles matrix $\tilde{X}_{n\times q}$

$$\tilde{x}_i = (\frac{k_i^1}{p}, ..., \frac{k_i^q}{p})$$

The inertia criterion

Transforming $X_{n\times p}$ (categorical) in $\tilde{X}_{n\times q}$ (numerical)

- $X_{n\times p} \rightarrow$ indicator matrix $\mathcal{K}_{n\times q}$ with 0 or 1 entries
- $\mathcal{K}_{n\times q}$ contingency table $\rightarrow$ row profiles matrix $\tilde{X}_{n\times q}$

$$\tilde{x}_i = (\frac{k_i^1}{p}, ..., \frac{k_i^q}{p})$$

Calculating the inertia on the rows of $\tilde{X}$:

- row-profiles weights : $(\frac{1}{n}, ..., \frac{1}{n})$
- the χ^2-distance : the Euclidean distance ($M = I$) weighted by the inverse of the column-profiles weights $(\frac{k_{.1}}{np}, ..., \frac{k_{.q}}{np})$

PART 3

DIVCLUS-T

DIVCLUS-T

First : how to divide a cluster C_l ?

- Find a bi-partition $(A, \bar{A})$ of C_l of smallest within-cluster inertia:

DIVCLUS-T

First : how to divide a cluster C_l ?

- Find a bi-partition $(A, \bar{A})$ of C_l of smallest within-cluster inertia:

  $\Rightarrow$ maximizing the between-cluster inertia:

  $$B(A, \bar{A}) = \frac{\mu_A \mu_{\bar{A}}}{\mu_A + \mu_{\bar{A}}} d^2(g_A, g_{\bar{A}}) = I(C_l) - \underbrace{I(A) + I(\bar{A})}_{W(A,\bar{A})}$$

  $\Rightarrow$ maximizing the inertia variation

DIVCLUS-T

First : how to divide a cluster C_l ?

- Find a bi-partition $(A, \bar{A})$ of C_l of smallest within-cluster inertia:

  $\Rightarrow$ maximizing the between-cluster inertia:

  $$B(A,\bar{A}) = \frac{\mu_A \mu_{\bar{A}}}{\mu_A + \mu_{\bar{A}}}\, d^2(g_A, g_{\bar{A}}) = I(C_l) - \underbrace{I(A) + I(\bar{A})}_{W(A,\bar{A})}$$

  $\Rightarrow$ maximizing the inertia variation

- $2^{n_l - 1} - 1$ possible bi-partitions

DIVCLUS-T

First : how to divide a cluster C_l ?

- Find a bi-partition $(A, \bar{A})$ of C_l of smallest within-cluster inertia:

  $\Rightarrow$ maximizing the between-cluster inertia:

  $$B(A,\bar{A}) = \frac{\mu_A \mu_{\bar{A}}}{\mu_A + \mu_{\bar{A}}} d^2(g_A, g_{\bar{A}}) = I(C_l) - \underbrace{I(A) + I(\bar{A})}_{W(A,\bar{A})}$$

  $\Rightarrow$ maximizing the inertia variation

- $2^{n_l - 1} - 1$ possible bi-partitions

  $\Rightarrow$ CART monothetic approach is used to reduce the number of bi-partitions to evaluate. The within-cluster inertia criterion is minimized among the set of bi-partitions induced by the set of all possible binary questions

DIVCLUS-T

The “dogs” dataset example

$\Rightarrow$ 2 binary variables $\rightarrow$ two bi-partitions
5 variables with 3 categories $\rightarrow 5 \times 3$ bi-partitions

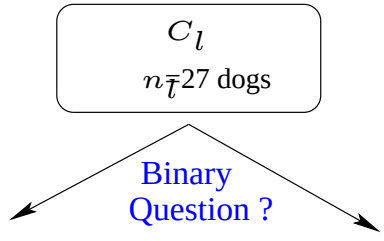
$\Rightarrow$ only 17 binary questions and bi-partitions to be evaluated

$\Rightarrow$ the question “Is the size large?” gives the smallest within-cluster inertia

$\Rightarrow$ the bi-partition is:

15 “large” dogs

12“small or medium” dogs



DIVCLUS-T

The “dogs” dataset example

$\Rightarrow$ 2 binary variables $\rightarrow$ two bi-partitions
5 variables with 3 categories $\rightarrow 5 \times 3$ bi-partitions

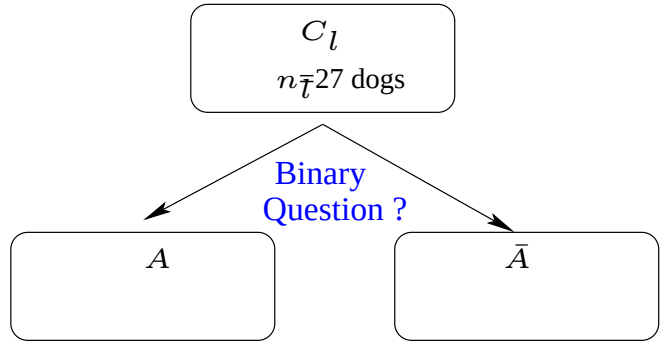
$\Rightarrow$ only 17 binary questions and bi-partitions to be evaluated

$\Rightarrow$ the question “Is the size large?” gives the smallest within-cluster inertia

$\Rightarrow$ the bi-partition is:

15 “large” dogs

12“small or medium” dogs



DIVCLUS-T

The “dogs” dataset example

$\Rightarrow$ 2 binary variables $\rightarrow$ two bi-partitions
5 variables with 3 categories $\rightarrow 5 \times 3$ bi-partitions

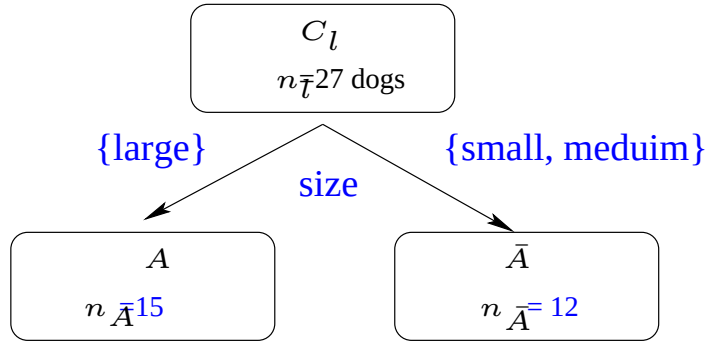
$\Rightarrow$ only 17 binary questions and bi-partitions to be evaluated

$\Rightarrow$ the question “Is the size large?” gives the smallest within-cluster inertia

$\Rightarrow$ the bi-partition is:

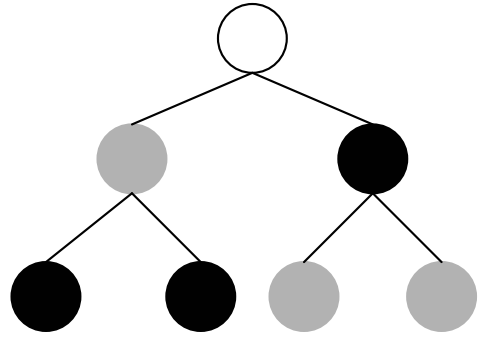
15 “large” dogs

12“small or medium” dogs



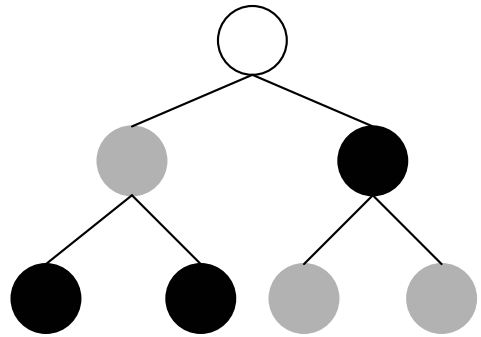
DIVCLUS-T

Second: how to choose the cluster $C_l = A \cup \bar{A}$ to split ?



DIVCLUS-T

Second: how to choose the cluster $C_l = A \cup \bar{A}$ to split ?



- Choose the cluster such that the $(k+1)$-clusters partition has the smallest within-cluster inertia

  $\Rightarrow$ Because $P_{k+1} = P_k \cup \{A, \bar{A}\} - \{C_l\}$

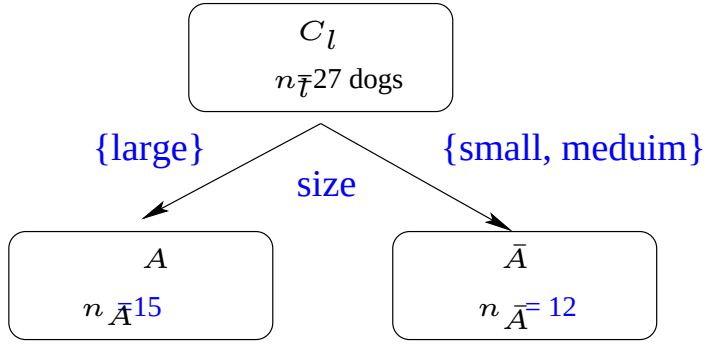
$$W(P_{k+1}) = W(P_k) - I(C_\ell) + I(A) + I(\bar{A})$$

  $\Rightarrow$ choice of C_l which maximizes the inertia variation:

$$B(A, \bar{A}) = I(C_l) - I(A) + I(\bar{A})$$

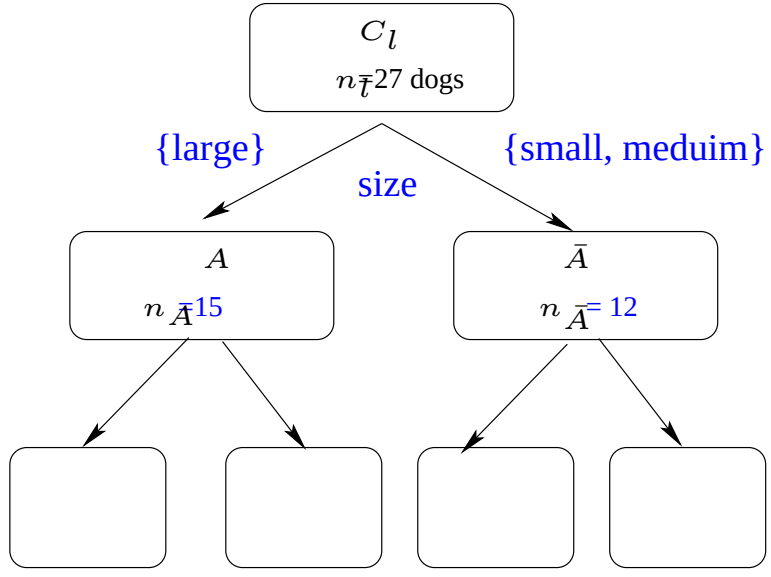
DIVCLUS-T

The “dogs” dataset example



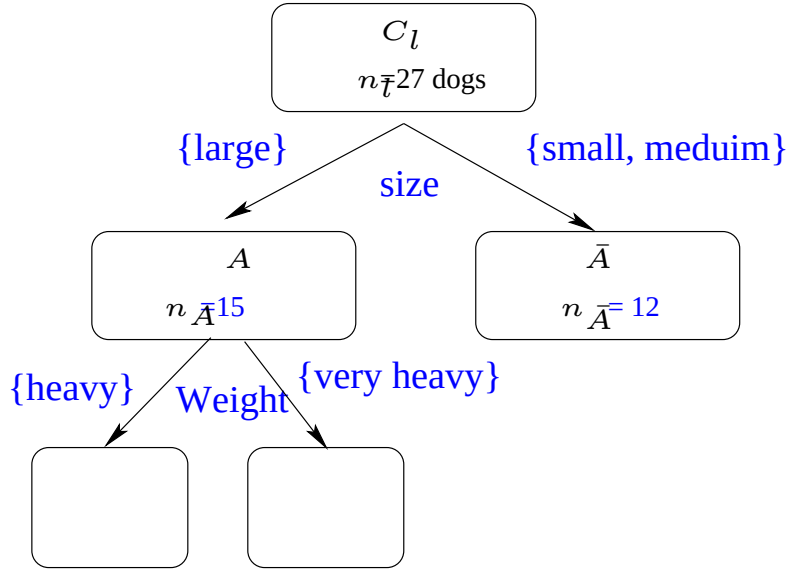
DIVCLUS-T

The “dogs” dataset example



DIVCLUS-T

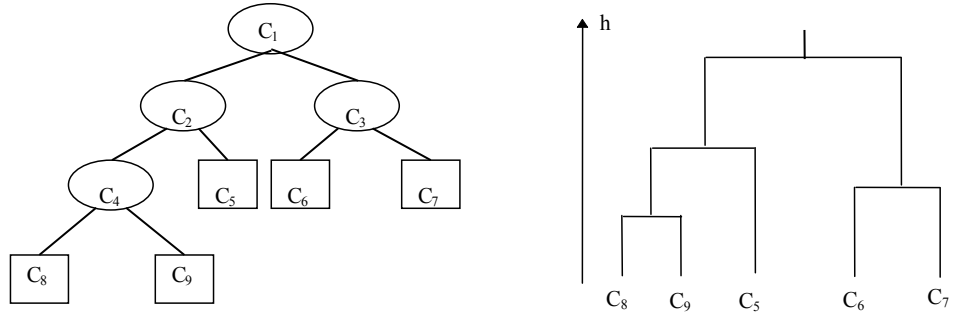
The “dogs” dataset example



DIVCLUS-T

Third: how to defined the hierarchical level ?

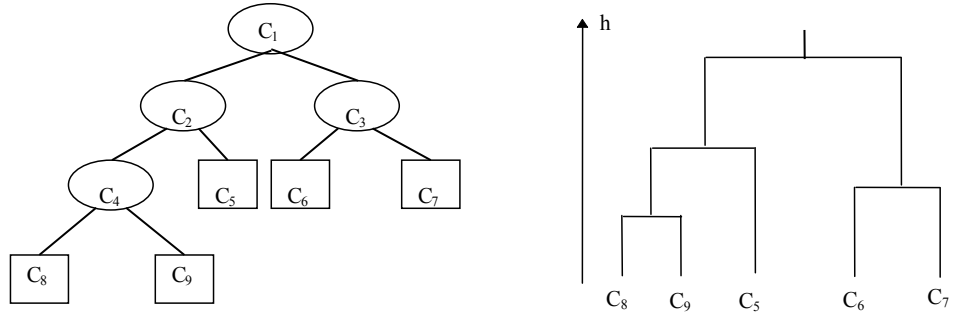
- The number of divisions is fixed



DIVCLUS-T

Third: how to defined the hierarchical level ?

- The number of divisions is fixed



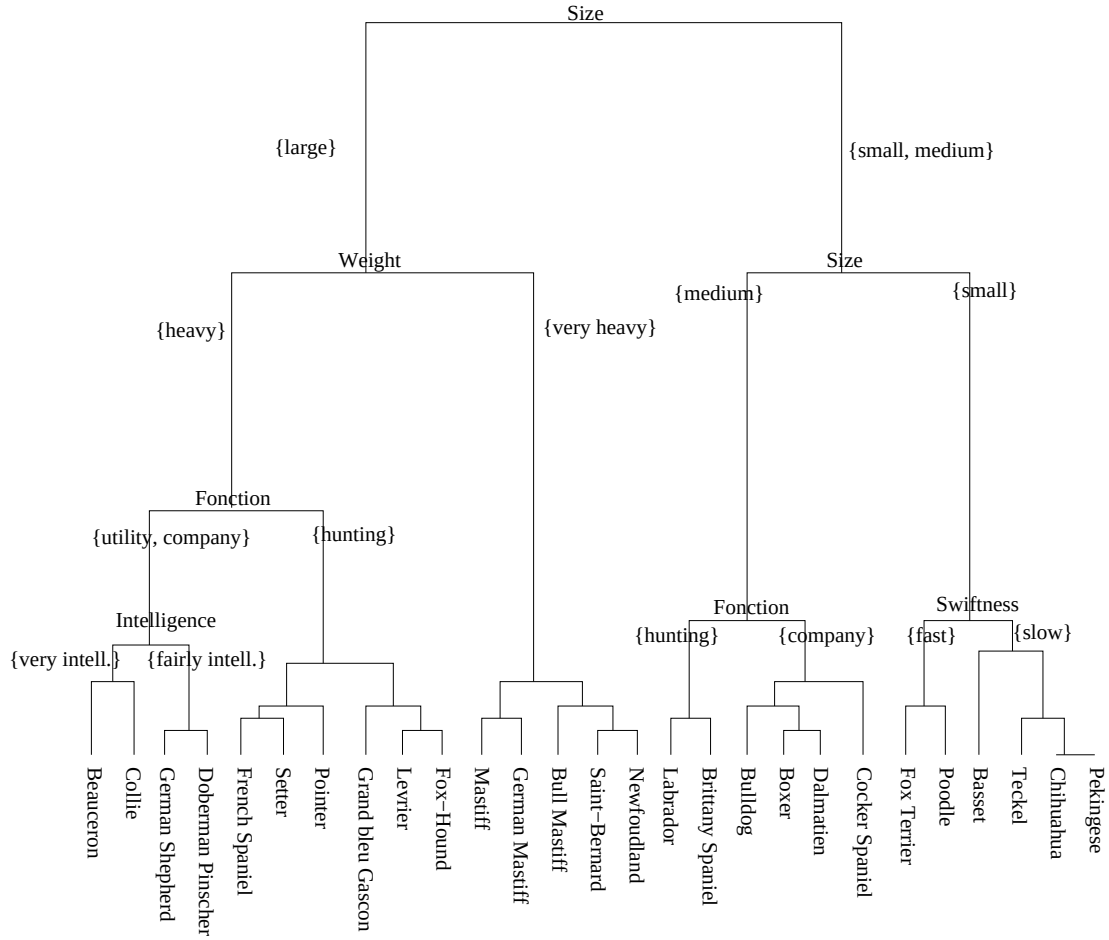
- The hierarchical level is the inertia variation:

$$h(C_\ell) = B(A, \bar{A}) = \frac{\mu_A \, \mu_{\bar{A}}}{\mu_A + \mu_{\bar{A}}} \, d^2(g_A, g_{\bar{A}})$$

$\Rightarrow$ same hierarchical level than for the Ward hierarchy

DIVCLUS-T

The “dogs” dataset example



Remark: the dendrogram obtained with Ward on all the components from Multiple Correspondance Analy-

PART 3

Comparison with Ward and the k-means

Empirical comparison

QUESTION: What is the price paied, in term of inertia, for this additional monothetic interpretation ?

Six data sets of the UCI Machine Learning repository

(*http://www.ics.uci.edu/ mlearn/MLRepository.html*)

Name	Type	Nb objects	Nb variables(nb categories)
Glass	numerical	214	8
Pima Indians diabete	numerical	768	8
Abalone	numerical	4177	7
Zoo	categorical	101	15(2) + 1(6)
Solar Flare	categorical	323	2(6) + 1(4) + 1(3) + 6(2)
Contraceptive Method Choice (CMC)	categorical	1473	9(4)

$\Rightarrow$ DIVCLUS-T, WARD and the k-means, three methods based on the minimization if the within-cluster inertia, have been applied on the six datasets

$\Rightarrow$ for the partitions from 2 to 15 clusters obtained with the three methods, the percentage of explained inertia has been calculated

Empirical comparison

The percentage E of inertia explained by a partition P is:

$$E(P) = (1 - \frac{W(P)}{I(O)}) \times 100$$

$\Rightarrow$ Measures the part of the inertia of O explained by P

$\Rightarrow$ Minimizing the within-cluster inertia is equivalent to maximize the explained inertia

$\Rightarrow$ $E(P) \nearrow$ when the number of clusters $\nearrow$ and $E(P)$ equals:

0 for the one-cluster partition

100 for the n-clusters (singleton) partition

Empirical comparison

The percentage E of inertia explained by a partition P is:

$$E(P) = (1 - \frac{W(P)}{I(O)}) \times 100$$

$\Rightarrow$ Measures the part of the inertia of O explained by P

$\Rightarrow$ Minimizing the within-cluster inertia is equivalent to maximize the explained inertia

$\Rightarrow$ $E(P) \nearrow$ when the number of clusters $\nearrow$ and $E(P)$ equals:

0 for the one-cluster partition

100 for the n-clusters (singleton) partition

Comparing two partitions

$\Rightarrow$ P is better than P' if $E(P) > E(P')$

$\Rightarrow$ P and P' must have the same number of clusters

Empirical comparison

Percentage of the explained inertia obtained with the three numerical datasets and the three clustering methods

	Glass			Pima			Abalone		
K	DIV	WARD	W+km	DIV	WARD	W+km	DIV	WARD	W+km
2	21.5	22.5	22.8	14.8	13.3	16.4	60.2	57.7	60.9
3	33.6	34.1	34.4	23.2	21.6	24.5	72.5	74.8	76.0
4	45.2	43.3	46.6	29.4	29.4	36.2	81.7	80.0	82.5
5	53.4	53.0	54.8	34.6	34.9	40.9	84.2	85.0	86.0
6	58.2	58.4	60.0	38.2	40.0	45.3	86.3	86.8	87.8
7	63.1	63.5	65.7	40.9	44.4	48.8	88.3	88.4	89.6
8	66.3	66.8	68.9	43.2	47.0	51.1	89.8	89.9	90.7
9	69.2	69.2	71.6	45.2	49.1	52.4	91.0	90.9	91.7
10	71.4	71.5	73.9	47.2	50.7	54.1	91.7	91.6	92.4
11	73.2	73.8	75.6	48.8	52.4	56.0	92.0	92.1	92.8
12	74.7	76.0	77.0	50.4	53.9	58.0	92.3	92.4	93.0
13	76.2	77.6	78.7	52.0	55.2	58.8	92.6	92.7	93.3
14	77.4	79.1	80.2	53.4	56.5	60.0	92.8	93.0	93.7
15	78.5	80.4	81.0	54.6	57.7	61.0	93.0	93.2	93.9

Empirical comparison

Three numerical datasets

	Glass			Pima			Abalone		
K	DIV	WARD	W+km	DIV	WARD	W+km	DIV	WARD	W+km
2	21.5	22.5	22.8	14.8	13.3	16.4	60.2	57.7	60.9
3	33.6	34.1	34.4	23.2	21.6	24.5	72.5	74.8	76.0
4	45.2	43.3	46.6	29.4	29.4	36.2	81.7	80.0	82.5
5	53.4	53.0	54.8	34.6	34.9	40.9	84.2	85.0	86.0
6	58.2	58.4	60.0	38.2	40.0	45.3	86.3	86.8	87.8
7	63.1	63.5	65.7	40.9	44.4	48.8	88.3	88.4	89.6
8	66.3	66.8	68.9	43.2	47.0	51.1	89.8	89.9	90.7
9	69.2	69.2	71.6	45.2	49.1	52.4	91.0	90.9	91.7
10	71.4	71.5	73.9	47.2	50.7	54.1	91.7	91.6	92.4
11	73.2	73.8	75.6	48.8	52.4	56.0	92.0	92.1	92.8
12	74.7	76.0	77.0	50.4	53.9	58.0	92.3	92.4	93.0
13	76.2	77.6	78.7	52.0	55.2	58.8	92.6	92.7	93.3
14	77.4	79.1	80.2	53.4	56.5	60.0	92.8	93.0	93.7
15	78.5	80.4	81.0	54.6	57.7	61.0	93.0	93.2	93.9

- Glass: WARD and DIVCLUS-T perform more or less similarly
- Pima : DIVCLUS-T $\geq$ WARD until 4 clusters and then WARD $\geq$ DIVCLUS-T
- Abalone : DIVCLUS-T $\geq$ WARD until 4 clusters and then WARD$\sim$ DIVCLUS-T

Empirical comparison

DIVCLUS-T **seems** to perform better

$\Rightarrow$ for few clusters partitions (maby because the few clusters partitions are found in the first steps of DIVCLUS-T whereas they are found in the last steps of WARD)

$\Rightarrow$ for bigger datasets (maby because the number of bi-partitions induced by numerical binary questions depends on the number of objects)

Empirical comparison

Percentage of explained inertia for the three categorical datasets

	Zoo			Solar Flare			CMC		
K	DIV	WARD	W+km	DIV	WARD	W+km	DIV	WARD	W+km
2	23.7	24.7	26.2	12.7	12.6	12.7	8.4	8.2	8.5
3	38.2	40.8	41.8	23.8	22.4	23.8	14.0	13.1	14.8
4	50.1	53.7	54.9	32.8	29.3	33.1	18.9	17.3	20.5
5	55.6	60.4	61.0	38.2	35.1	38.4	23.0	21.3	24.0
6	60.9	64.3	65.1	43.0	40.0	42.7	26.3	24.9	27.7
7	65.6	67.5	68.4	47.7	45.0	47.6	28.4	28.1	29.8
8	68.9	70.6	71.3	51.6	49.8	52.1	30.3	30.7	32.7
9	71.8	73.7	73.7	54.3	53.5	54.6	32.1	33.4	35.2
10	74.7	75.9	75.9	57.0	57.1	58.3	33.8	35.5	37.7
11	76.7	77.5	77.5	59.3	60.4	61.7	35.5	37.5	40.1
12	78.4	79.1	79.1	61.3	62.9	64.4	36.9	39.4	41.5
13	80.1	80.6	80.6	63.1	65.2	65.7	38.1	41.0	42.9
14	81.3	81.8	81.8	64.5	66.2	67.7	39.2	42.0	44.2
15	82.8	82.8	82.8	65.8	68.6	69.3	40.3	43.1	44.9

- Zoo : WARD $\geq$ DIVCLUS-T
- Solar Flare : DIVCLUS-T $\geq$ WARD until 10 clusters
- CMC : DIVCLUS-T $\geq$ WARD until 8 clusters

Computational complexity

For numerical data :

- DIVCLUS-T: $o(Kpn(log(n) + p))$
- WARD: $o(pn^2) \rightarrow$ DIVCLUS-T complexity is better for small values of K
- K-means: $o(KpnT) \rightarrow$ DIVCLUS-T complexity is better when $log(n) + p < T$

Conclusion

- DIVCLUS-T is a descendant hierarchical clustering method:
  - Advantage: the clusters have by construction natural interpretations
  - Question: What is the price paied, in term of inertia, for this additional monothetic interpretation ?
- Comparison with WARD and the k-means
  - The first results on the 6 datasets show “comparable” results in term of inertia particulary for the few clusters partitions. Further study remain necessary
- Computational complexity
  - DIVCLUS-T is efficient for numerical data and for categorical data...

Conclusion

If the user is interested in ratherly large partitions for instance in order to reduce the number of objects, WARD and the k-means are certainly more effecient than DIVCLUS-T.
But if the user is interested in few clusters partitions with good interpretations, DIVCLUS-T seems to be an interesting alternative to classical methods.

Empirical comparison of a monothetic divisive clustering method with the Ward and the k-means clustering methods

Marie Chavent, Yves Lechevallier

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,

Université Bordeaux1, 351, Cours de la libération, 33405 Talence Cedex, France

chavent@math.u-bordeaux1.fr

Institut National de Recherche en Informatique et en Automatique,

Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France

Yves.Lechevallier@inria.fr