Logical characterization of groups of data: a comparative study

Arthur Chambon, Tristan Boureau, Frédéric Lardeux, Frédéric Saubion

Noname manuscript No.
(will be inserted by the editor)

Logical Characterization of Groups Data: a Comparative Study

Arthur Chambon · Tristan Boureau · Frédéric Lardeux · Frédéric Saubion

the date of receipt and acceptance should be inserted later

Abstract This paper presents an approach for characterizing groups of data represented by Boolean vectors. The purpose is to find minimal set of attributes that allow to distinguish data from different groups. In this work, we precisely defined the multiple characterization problem and the algorithms that can be used to solve its different variants. Our data characterization approach can be related to Logical Analysis of Data and we propose thus a comparison between these two methodologies. The purpose of this paper is also to precisely study the properties of the solutions that are computed with regards to the topological properties of the instances. Experiments are thus conducted on real biological data.

Keywords Logical analysis of data · Characterization of multiple groups of data

1 Introduction

Let us consider a set of observations Ω, whose elements are data expressed over a set of Boolean attributes $\mathcal{A}$. Given a partition of Ω into several groups -subsets- of data, the Multiple Characterization Problem (MCP) consists in finding a subset of attributes that discriminates these groups. A solution of this problem is thus a reduction of the initial attributes set such that there does not exist two identical observations in two different groups, with regards to these selected attributes. This general problem has a lot of extensions like the $Min - MCP$ which computes a minimal solution in terms of number of attributes. MCP has many applications, especially in the processing of biological data, where observations of individuals (e.g., patients, plants, animals...) can be studied by means of groups according to their characteristics (behavior, pathology, common phenotypic/genotypic properties...). Such observations may correspond to aggregation of biological attributes

A. Chambon · F. Lardeux · F. Saubion
LERIA, University of Angers (France)
E-mail: {firstname.lastname}@univ-angers.fr

T.Boureau
IRHS, University of Angers (France)
E-mail: tristan.boureau@univ-angers.fr

(e.g., presence or absence of genes or biological markers...). Therefore, the purpose of MCP can be interpreted as the extraction of a signature that is sufficient to explain *a priori* the membership of observations to groups initially provided by the user/expert. This signature may also be used in order to assign groups to new incoming observations.

Example 1 Let us consider the following instance of MCP, with 7 observations, 3 groups and 8 attributes ($[a \dots h]$).

Observations	Groups	Attributes							
		a	b	c	d	e	f	g	h
1	1	0	1	0	1	0	1	1	0
2		1	1	0	1	1	0	0	1
3		0	1	1	0	1	0	0	1
4	2	1	0	1	0	1	0	1	1
5		0	0	0	1	1	1	0	0
6	3	1	1	0	1	0	1	0	1
7		0	0	1	0	1	0	1	0

The subset of attributes $\{f, g, h\}$ can be used to discriminate the three groups. Focusing on attributes $\{f, g, h\}$, no identical lines appear in two different groups (even if the same line appears twice in the same group, for instance in Group 1). Turning now to the interpretation of the solution, one may observe that the first group can be characterized, for instance, by the following Boolean formula $(f \wedge g) \vee (\bar{f} \wedge \bar{g})$, if attributes are seen as Boolean variables. This formula constitutes a more compact representation of some informations contained in Group 1. This interpretation corresponds to the restriction of the initial problem to the characterization of Group 1 against the other groups.

Therefore, the initial table (matrix) that represents the observations according to the attributes, can be interpreted as a partially defined Boolean function that has to be minimized. Of course, this consideration is only valid when considering the characterization of one group against the others. In this case, this minimization problem has been shown to be NP-hard [22]. When considering only two groups - for instance a group of positive observations and a group of negative observation, which is a very common problem in machine learning - the concept of Logical Analysis of Data has been proposed in [8].

Related works

The feature selection problem [11,20] aims at selecting a subset of attributes (features) that may efficiently describe the data. Therefore, our approach can be related to this problem. Many different methods are available for feature selection, for instance based on statistical evaluation of the relevance of the attributes. From the machine learning point of view it is also important to assess the ability to generalize to new incoming data. Feature selection methods are useful in order to improve classifiers such as support vector machines. Note that feature selection techniques are also relevant for data visualization and data compression. Here our problem is different since we are mainly interested in finding attributes-based explanations for given groups of data built by experts. No *a posteriori* validation

is considered. In fact, the selected attributes ensure that the groups of data can be identified by a logical combination of these attributes. We have indeed tested some statistical feature selection methods on real data from biology [9] in order to filter the initial set of attributes, without obtaining satisfactory results with regards to our problem. The MCP solutions are mainly composed of attributes with low scores with regards to the tested feature selection methods. We also observe that attributes with highest scores do not participate to solutions. It suggests that MCP solutions are not built with the attributes that are the most correlated to the groups.

The Logical Analysis of Data (LAD) is a methodology that is mainly based on the notion of pattern. Considering an instance constituted by two groups of Boolean observations (called positive an negative observations), the LAD method consists in finding a subset of attributes that have the similar values on some observations of the positive set while these values cannot be observed in the negative group. This subset is called a pattern and the purpose is to find patterns that are shared by as many positive observations as possible. Our approach differs from LAD on several aspects:

- we simultaneously consider several groups to be mutually discriminated,
- the objective is to provide a formula that minimizes the number of used attributes rather than patterns that covers observations.

Nevertheless, our proposal is clearly related to the general of LAD.

Examples of applications of MCP

Characterization of multiple groups of Boolean data has been applied in plant biology in order to identify strains of the species *Xanthomonas axonopodis* [9], a family of bacteria that cause different diseases on many plants. In this context, groups consist of bacterial strains that share the same pathogenic behavior. Each strain is identified by the presence or absence (1 or 0) of genes. Note that each strain has a narrow range of potential hosts: for instance, genes that are necessary to infect Bean are shared by all strains infecting Bean and another combination of genes necessary to infect Tangerine is shared by all strains infecting Tangerine. Similar examples may be found for bacterial pathogens on animals: strains of *Salmonella typhi* are pathogenic on humans whereas strains of *Salmonella enteritidis* are pathogenic on chickens. Numerous genomes of pathogenic bacteria are available in public databases, and presence or absence of genes can be expressed into Boolean matrices.

This approach can also be used to identify biological markers that are useful in order to predict the evolution of specific diseases on patients. In this paper, we use biological data from patients that suffer from acute leukemia. The purpose is to identify combination of mutated genes that can be used to predict if the patients belong to a remission or to a relapse group and need a specific treatment.

Contributions of the paper

- Based on previous preliminary works [12,10], we propose here a new formulation of the multiple characterization problem that allows us to clearly define and study the properties of the instances as well as the properties of the solutions that can be computed.

- We propose a comparison with LAD methodology and compare how the solutions provided by these two approaches could be related.
- Based on a set of real biological instances, we provide an experimental evaluation of our approach using topological measures (i.e., Boolean distance based). In particular we establish that classic clustering techniques are not efficient to characterize those real data, since the groups formed by experts do not exhibit classic similarity properties. Indeed, our method that compute combinations of attributes instead of similarity based aggregations of attributes allows the practitioners to discover alternative relationships between attributes within her groups of data.

Organization of the paper

In Section 2, we recall the main concepts related to Logical Analysis of Data. In Section 3 and 4 the multiple characterization problem is precisely formulated, using two different possible formalisms. Its properties are studied in Section 5. Solving algorithms are presented in Section 6. A comparison between the Logical Analysis of Data methodology and our approach is proposed in Section 7. The experimental setup is presented in Section 8, while results and analysis are described in Section 9.

2 Logical Analysis of Data

Logical Analysis of Data (LAD) [5,7,8,13,16] is a data analysis method that aims to find some patterns playing an important role in the classification of data. Contrary to statistical data analysis techniques, LAD is mainly based on combinatorial optimization and logics and, more specifically, on the concept of partially defined Boolean functions. Hence, as presented in the introduction, LAD considers two groups of observations, represented as Boolean vectors in an attributes space. LAD has been applied to many domains: biology and medicine [19,3,4], engineering [6], transportation [15].

Based on [16] and [7], we recall the main concepts related to LAD and more especially, the notion of partially defined Boolean function.

Definition 1 A Boolean function of n variables, $n \in \mathbb{N}$, is a mapping $\mathbb{B}^n \mapsto \mathbb{B}$, where $\mathbb{B}$ is the set $\{0, 1\}$.

Definition 2 A vector $x \in \mathbb{B}^n$ is a *true vector* (resp. *false vector*) of the Boolean function f if $f(x) = 1$ (resp. $f(x) = 0$). $T(f)$ (resp. $F(f)$) is the set of *true vectors* (resp. *false vectors*) of a Boolean function f.

Definition 3 A partially defined Boolean function (pdBf) on $\mathbb{B}^n$ is a pair (P,N) such that $P,N \subseteq \mathbb{B}^n$ and $P \cap N = \emptyset$.

P is thus the set of positive vectors, and N the set of negative vectors of the pdBf (P,N). Let us remark that the condition $P \cap N = \emptyset$ may not be satisfied in certain real world data sets of classification problems.

The notion of partially defined Boolean function is generalized by the following notion of term used in [16,7].

Definition 4 A term is a Boolean function t_{σ^+,σ^-} whose true set $T(t_{\sigma^+,\sigma^-})$ is of the form

$$T(t_{\sigma^+,\sigma^-}) = \{x \in \mathbb{B}^n | x_i = 1 \ \forall i \in \sigma^+ \text{ and } x_j = 0 \ \forall j \in \sigma^-\}$$

for some set $\sigma^+, \sigma^- \subseteq \{1, 2, ..., n\}$, $\sigma^+ \cap \sigma^- = \emptyset$.

A term t_{σ^+,σ^-} can be represented by an *elementary conjunction*, i.e. a Boolean expression of the form

$$t_{\sigma^+,\sigma^-}(x) = (\bigwedge_{i \in \sigma^+} x_i) \wedge (\bigwedge_{j \in \sigma^-} \bar{x}_j)$$

We recall now a key concept for LAD: the notion of pattern, whose aim is to identify a set of attributes that have identical values for several observations of the positive group P. These common values constitute a pattern that must not appear in the negative group N.

Definition 5 A pattern of a pdBf (P,N) is a term t_{σ^+,σ^-} such that $|P \cap T(t_{\sigma^+,\sigma^-})| > 0$ and $|N \cap T(t_{\sigma^+,\sigma^-})| = 0$.

Definition 6 Given a term t, $Var(t_{\sigma^+,\sigma^-})$ is the set of variables defining the term $(Var(t_{\sigma^+,\sigma^-}) = \{x_i | i \in \sigma^+ \cup \sigma^-\})$.

Given a pattern p of a pdBf (P,N), the set $P \cap T(p)$ is said to be covered by the pattern p.

One goal of LAD is to compute optimal patterns $p_i = argmax_{p_i}(|P \cap T(p_i)|)$, i.e., patterns that cover as many observations as possible. This approach may be related to data mining concepts such as frequent itemset identification [2].

Example 2 Let us consider the instance in Example 1, where Group 1 is the set of positive observations and Group 2 and Group 3 are merged into the set of negative observations.

Observations	Groups	Attributes							
		a	b	c	d	e	f	g	h
1	P	0	1	0	1	0	1	1	0
2		1	1	0	1	1	0	0	1
3		0	1	1	0	1	0	0	1
4	N	1	0	1	0	1	0	1	1
5		0	0	0	1	1	1	0	0
6		1	1	0	1	0	1	0	1
7		0	0	1	0	1	0	1	0

$p_1 = \bar{a} \wedge b$ and $p_2 = \bar{f} \wedge \bar{g}$ are 2 patterns covering observations 1 and 3 for p_1 and 2 and 3 for p_2.

In Example 2, we note that $p_2 = \bar{f} \wedge \bar{g}$ and $p_3 = f \wedge g$ are 2 patterns using identical attributes $(\sigma^+_{p_2} \cup \sigma^-_{p_2} = \sigma^+_{p_3} \cup \sigma^-_{p_3})$.

3 The Multiple Characterization Problem

The purpose of this section is to formalize the Multiple Characterization Problem (MCP) described in the introduction.

3.1 Problem Definition

Definition 7 (MCP instance) An instance of the MCP is a tuple $(\Omega, \mathcal{A}, D, G)$ defined by a set of observations Ω, whose elements are data expressed over a set of Boolean attributes $\mathcal{A}$ encoded by a Boolean matrix of data $D_{|\Omega|\times|\mathcal{A}|}$ and a function $G : \Omega \to \mathbb{N}$, such that $G(o)$ is the group assigned to the observation $o \in \Omega$.

The data matrix D is defined as follows:

- the value $D[o, a]$ represents the presence/absence of the attribute a in the observation o.
- a line $D[o, .]$ represents thus the Boolean vector of presence/absence of the different attributes in the observation o.
- a column $D[., a]$ represents thus the Boolean vector of presence/absence of the attribute a in all the observations.

In the following we are only interested in satisfiable MCP, i.e. such that D does not contain two identical observations in two different groups (see [12]).

Property 1 A MCP instance $(\Omega, \mathcal{A}, D, G)$ is satisfiable iff:

$\nexists (o, o') \in \Omega^2$ such that $D[o, .] = D[o', .]$ and $G(o) \neq G(o')$

D^A is the data matrix reduced to the subset of attributes $A \subset \mathcal{A}$. Given a satisfiable instance $(\Omega, \mathcal{A}, D, G)$, $D_{\succ}$ is the matrix D where identical lines have been deleted (even if two identical lines correspond indeed to two different observations on real data, we reduce the matrix for computational purpose).

Example 3 Consider the instance $(\Omega, \mathcal{A}, D, G)$ from Example 2. Consider the matrix $D^{\{f,g\}}$ reduced to the subset of attributes $\{f, g\}$.

We have[1] $D^{\{f,g\}}=$

Observations	Groups	Attributes	
		f	g
1	P	1	1
2		0	0
3		0	0
4	N	0	1
5		1	0
6		1	0
7		0	1

Some observations coming from the same group are now identical. We can reduce the matrix by considering only the observations in $\Omega_{\{f,g\}} = \{1, 2, 4, 5\}$.

Thus, we have $D_{\succ}^{\{f,g\}}=$

Observations	Groups	Attributes	
		f	g
1	P	1	1
2		0	0
4	N	0	1
5		1	0

[1] Note that for simplicity, we present the full array that contains the data matrix (which corresponds thus only the Boolean part of the array).

3.2 Solutions

Solving an instance $(\Omega, \mathcal{A}, D, G)$ consists in finding a subset of attributes $S \subseteq \mathcal{A}$ such that two observations from two different groups are always different on at least one attribute in S (i.e., $D^S_{\succ}$ has no identical line). According to our previous definition of satisfiability (Definition 1), it consists in finding a set S such that (Ω, S, D^S, G) is satisfiable.

Definition 8 Given an instance $(\Omega, \mathcal{A}, D, G)$, a subset of attributes $S \subseteq \mathcal{A}$ is a solution iff $\forall (o, o') \in \Omega^2, G(o) \neq G(o') \rightarrow D^S[o,.] \neq D^S[o',.]$. In this case, the matrix D^S is called a solution matrix.

An instance of the MCP may have several solutions of different sizes. It is therefore important to define an ordering on solutions in order to compare and classify them. In particular, for a given solution S, adding an attribute generates a new solution $S' \supset S$. In this case we say that S' is dominated by S.

Definition 9 A solution S is non-dominated iff $\forall s \in S$, $\exists (o, o') \in \Omega^2$ such that $G(o) \neq G(o')$ and $D^{S\setminus\{s\}}[o,.] = D^{S\setminus\{s\}}[o',.]$.

Among these solutions, we are interested in computing solutions of minimal size with regards to the attributes they involve.

Definition 10 A solution S is minimal iff $\nexists S'$ with $|S'| < |S|$ s.t. S' is a solution.

According to our notion of dominance between solutions, a minimal solution in not dominated by any other solutions.

Intuitively, a minimal (non dominated) solution cannot be reduced unless two identical lines appear in two different groups (and consequently the reduced set of attributes is not a solution).

According to these notions of solutions, given an MCP instance $\mathcal{I} = (\Omega, \mathcal{A}, D, G)$, we may identify thus several problems :

- $Sol - MCP$: computing a solution of $\mathcal{I}$ (according to Definition 8)
- $Min - MCP$: computing a minimal solution of $\mathcal{I}$ (according to Definition 10)
- $1Min - MCP$: computing a minimal solution for 1 group against the others (in order to characterize a given group).
- $NonDomAll - MCP$ computing all non dominated solutions of $\mathcal{I}$ (According to Definition 9)
- $MinAll - MCP$ computing all minimal solutions of $\mathcal{I}$ (According to Definition 9)

In the following, according to practitioners requirements, we are mainly interested in $MinAll - MCP$.

We propose to recall the main concepts and notations used in LAD and MCP in Table 1.

4 Reformulation of the MCP

In the remaining of this paper, we use the classical Boolean notations: $\wedge$ is the conjunction, $\vee$ is the disjunction, $\oplus$ is the exclusive disjunction Xor and $\odot$ is the XNor logical operator ($[0,0,1,1] \odot [0,1,0,1] = [1,0,0,1]$).

Table 1 Notations

Notation	Description
LAD	Logical Analysis of Data
MCP	Multiple Characterization Problem
pdBf	Partially defined Boolean function
Ω	Set of observations (Boolean vectors)
$\mathcal{A}$	Set of Boolean attributes
P	Group of positive observations
N	Group of negative observations
G	Group function that assigns groups to observations
D	Data matrix (possibly used with different superscripts and subscripts)
C	Constraint matrix (possibly used with different superscripts and subscripts)
sim	similarity function on observations (variants with different subscripts)
$diff$	difference function on observations (variants with different subscripts)

4.1 Converting Characterization Requirements into Constraints

According to previous works on MCP [12], the minimum multiple characterization problem can be formulated as a linear program. Given an instance $(\Omega, \mathcal{A}, D, G)$, let us consider the following 0/1 linear program.

$$
\begin{aligned}
&min : \sum_{i=1}^{|\mathcal{A}|} y_i \\
&s.t. : \\
&C \,.\, Y^t \geqslant \mathbb{1}^t \\
&Y \in \{0,1\}^{|\mathcal{A}|}, Y = [y_1, ..., y_{|\mathcal{A}|}]
\end{aligned}
$$

where Y is a Boolean vector that encodes the presence/absence of the set of attributes in the solution (and Y^t is the transposed matrix of Y). C is a matrix that defines the constraints that must be satisfied in order to insure that Y is a solution. Let us denote Θ the set of all pairs $(o, o') \in \Omega^2$ such that $G(o) \neq G(o')$. For each pair of observations (o, o') that do not belong to the same group, defined by an element of Θ, one must insure that the value of at least one attribute differ from o to o'. This will be insured by the inequality constraint involving the $\mathbb{1}$ vector (here a vector of dimension $|\Theta|$ that contains only 1 values). The minimization objective function insures that we aim to find a minimal solution.

More formally, C is a Boolean matrix of size $|\Theta| \times |\mathcal{A}|$ constructed as follows:

- Each line is numbered by a couple of observations $(o, o') \in \Omega^2$ such that $G(o) \neq G(o')$ $((o, o') \in \Theta)$.
- Each column represents an attribute.
- $C[(o, o'), a] = 1$ if $D[o, c] \neq D[o', c]$, $C[(o, o'), a] = 0$ otherwise.
- We denote $C[(o, o'), .]$ the Boolean vector representing the differences between observations o and o' on each attribute. This Boolean vector is called constraint since one attribute a such $C[(o, o'), a] = 1$ must be selected in order to insure that no identical observations can be found in different groups .

Note that the constraints matrix C can be deduced from the data matrix D and, conversely, D can be deduced from C and a given vector $D_{[o,.]}$, $o \in \Omega$ (because C

assure for each value $D_{[o',i]}$ if this value is similar of different than the value $D_{[o,i]}$). As previously, C^A is the constraints matrix projected on the set of attributes $A \subset \mathcal{A}$. We also define a notion of reduction of the constraints matrix C.

Definition 11 A vector $C[(o,o'),.]$ is redundant if there exists a couple $(k,k') \in \Theta$ s.t. $(C[(o,o'),.] \wedge C[(k,k'),.]) = C[(k,k'),.]$.

Using this notion of redundancy, C can be thus reduced.

Definition 12 The reduced constraints matrix, denoted $C_{\succ}$, is the constraints matrix projected on the set Θ' where $\Theta' \subseteq \Theta$ the set of all pairs $(o,o') \in \Theta$ such that $C[(o,o'),.]$ is not redundant.

Of course, $C_{\succ}$ is of size $|\Theta'| \times |\mathcal{A}|$.

4.2 Solutions

The notion of solution can be redefined with regards to the new formulation of the problem. In the following, we denote $\mathbb{0}$ the vector that contains only 0 values.

Definition 13 Given a solution S and a constraint $C[(o,o'),.]$, S satisfies the constraint $C[(o,o'),.]$ if $\exists a \in S$ such that $C[(o,o'),a] = 1$ (i.e. $C^S[(o,o'),.] \neq \mathbb{0}$).

A solution is a set of attributes such that each constraint is satisfied, i.e., there is at least the value 1 in each line of C^S. We have thus the following obvious property.

Proposition 1 *A set of attributes $S \subseteq \mathcal{A}$ is a solution iff $\nexists (o,o') \in \Theta$ such that $C^S[(o,o'),.] = \mathbb{0}$.*

Let us consider the following example in order to illustrate the notions of solution and minimal solution in the context of constraints matrix.

Example 4 Let us consider the instance from Example 1. The matrix of constraints C associated with this matrix is:

Observations pairs	Attributes							
	a	b	c	d	e	f	g	h
(1,4)	1	1	1	1	1	1	0	1
(1,5)	0	1	0	0	1	0	1	0
(1,6)	1	0	0	0	0	0	1	1
(1,7)	0	1	1	1	1	1	0	0
(2,4)	0	1	1	1	0	0	1	0
(2,5)	1	1	0	0	0	1	0	1
(2,6)	0	0	0	0	1	1	0	0
(2,7)	1	1	1	1	0	0	1	1
(3,4)	1	1	0	0	0	0	1	0
(3,5)	0	1	1	1	0	1	0	1
(3,6)	1	0	1	1	1	1	0	0
(3,7)	0	1	0	0	0	0	1	1
(4,6)	0	1	1	1	1	1	1	0
(4,7)	1	0	0	0	0	0	0	1
(5,6)	1	1	0	0	1	0	0	1
(5,7)	0	0	1	1	0	1	1	0

$S = \{f, g, h\}$ is a non-dominated solution. If we compute the reduced constraints matrix $C_{\succ}^{S}$ projected on S, we obtain:

Observations pairs	Attributes		
	f	g	h
(1,5)	0	1	0
(1,7)	1	0	0
(4,7)	0	0	1

We may remark that a solution is non dominated if the reduced constraints matrix projected on this solution can be transformed by line permutation into a diagonal matrix (i.e., each constraint is satisfied by exactly one attribute). Here we can obtain the following diagonal matrix:

Observations pairs	Attributes		
	f	g	h
(1,7)	1	0	0
(1,5)	0	1	0
(4,7)	0	0	1

According to the previous remark, we can reformulate the concept of non-dominated solution and minimal solution from the constraints matrix.

Proposition 2 *A solution S is non-dominated iff $\forall s \in S$, $\exists (o, o') \in \Theta$ such that $C^{S \setminus \{s\}}[(o, o'), .] = \mathbb{0}$*

We have of course the same property concerning minimal solutions.

Proposition 3 *A solution S is minimal iff $\nexists S'$ with $|S'| < |S|$ such that $\forall (o, o') \in \Theta$, $C^{S'}[(o, o'), .] \neq \mathbb{0}$.*

5 Properties of the MCP

The purpose of this section is to exhibit properties and relationships between the two representations of the problem. Note that these properties are used in order to define our solving method.

5.1 Problem Properties

We describe now the properties of an instance $(\Omega, \mathcal{A}, D, G)$, and its associated constraint matrix C as defined above.

Proposition 4 *Given an instance $(\Omega, \mathcal{A}, D, G)$ with its constraint matrix C, a set of attributes $A \subseteq \mathcal{A}$ has one of the following properties:*

1. *A is not a solution $\Leftrightarrow C_{\succ}^{A} = \mathbb{0}$,*
2. *A is a non-dominated solution $\Leftrightarrow$ there exists a particular permutation φ on A such that $C_{\succ}^{\varphi(A)} = Id$ (where Id is the identical matrix),*
3. *A is a dominated solution otherwise.*

In Proposition 1, an instance of the MCP has a solution if and only if do not exist two identical observations from two different groups (i.e., $\nexists(o,o') \in \Theta$ s.t. $C^A[(o,o'),.] = \mathbb{0}$). Note that, if $\exists(o,o') \in \Theta$ s.t. $C^A[(o,o'),.] = \mathbb{0}$, then $C^A_{\succ} = \mathbb{0}$. Moreover, using Proposition 2, it is clear that if A is a solution, C^A contains one vector per attributes where only this attribute is set to 1. Of course, in $C^A_{\succ}$, only these vectors remain.

5.2 Reduction and Simplification of an Instance

In [12], it has been shown that if $D[.,a_1] = D[.,a_2]$ then it is possible to restrict the problem to $D^{\mathcal{A}\backslash\{a_2\}}$ or $D^{\mathcal{A}\backslash\{a_1\}}$. In [10], the concept of domination between attributes has been introduced. Given $(a,b) \in \mathcal{A}^2$, a dominates b, denoted $a \succ b$, iff $\forall S \subseteq \mathcal{A}\backslash\{b\}$, if $S \cup \{b\}$ is a non-dominated solution, then $S \cup \{a\}$ is also a non-dominated solution. Hence, we reformulate this property according to our new framework.

Proposition 5 *Given* $(a,b) \in \mathcal{A}^2$. *If* $(C[.,a] \vee C[.,b]) = C[.,a]$ *then* $a \succ b$.

Of course, all non-dominated solutions that contain at least one dominant attribute may be used to generate another solution by changing each dominated attribute by a non dominated attribute.

Given an instance $(\Omega, \mathcal{A}, D, G)$, let us consider $\mathcal{E} \subset \mathcal{A}$ the set of all dominated attributes. Then the MCP $(\Omega, \mathcal{A}\backslash\mathcal{E}, D', G)$ has at least one minimal solution in common with $(\Omega, \mathcal{A}, D, G)$.

5.3 Topological Properties of Instances

In this section, we define different measures in order to study later the properties of the observations and their distribution in the groups. Let us consider an instance $(\Omega, \mathcal{A}, D, G)$ and its constraints matrix C.

The following definition presents the notion of similarity between two observations.

Definition 14 Given $(o,o') \in \Omega^2$, $A \subseteq \mathcal{A}$, the similarity $sim(D^A,(o,o'))$ between o and o' is the mean of values of the vector $D^A_{[o,.]} \odot D^A_{[o',.]}$:

$$sim(D^A,(o,o')) = \frac{1}{|A|} \times \sum_{a \in A} (D^A[o,a] \odot D^A[o',a])$$

Using the previous definition, we may now define a notion of intragroup similarity for a group of a given MCP instance.

Definition 15 The intragroup similarity of a group is defined as the average of the similarities between the observations in this group.

For $A \subseteq \mathcal{A}$ and a group $G_g \subset \Omega$ s.t. $G_g = \{o | G(o) = g\}$,

$$sim_g(D^A) = \frac{1}{|G_g|!} \times \sum_{o,o' \in G_g} sim(D^A,(o,o'))$$

The intragroup similarity of an instance is defined as the average similarity between observations of the same group.

Definition 16 For $A \subseteq \mathcal{A}$,

$$sim_{intra}(D^A) = \frac{1}{|\Omega^2 \backslash \Theta|} \times \sum_{(o,o') \in \Omega^2 \backslash \Theta} sim(D^A, (o, o'))$$

The overall similarity of an instance is defined as the average of the similarities between all observations within the set of attributes $A = \mathcal{A}$.

Definition 17

$$sim_{over}(D^A) = \frac{1}{|\Omega^2|} \times \sum_{(o,o') \in \Omega^2} sim(D^A, (o, o'))$$

We want now to evaluate the difference between observations of different groups.

Definition 18 Given $(o, o') \in \Omega^2$, $A \in \mathcal{A}$, the difference between o and o', denoted $diff(D^a, (o, o'))$, is the mean value of the vector $D^A[o, .] \oplus D^A[o', .]$:

$$diff(D^A, (o, o')) = \frac{1}{|A|} \times \sum_{a \in A} (D^A[o, a] \oplus D^A[o', a])$$

Definition 19 The intergroup difference of an instance is defined as the average of the differences between observations from different groups.

For $A \in \mathcal{A}$,

$$diff_{inter}(D^A) = \frac{1}{|\Theta|} \times \sum_{(o,o') \in \Theta} diff(D^A, (o, o'))$$

We have the following property.

Proposition 6 *Given $A \in \mathcal{A}$,*

$$diff_{inter}(D^A) = \frac{1}{|\Theta|} \times \sum_{(i,j) \in \Theta} \sum_{c \in C} C^A_{[(i,j),c]}$$

Computing of the intergroup difference use actually the average number of differences between the observations of the set Θ.

6 Solving the MCP

Based on [10], we present algorithms in order to address the $NonDomAll - MCP$ and $MinAll - MCP$ problems.

6.1 Computation of all Non-dominated Solutions

Algorithm 1 aims at computing the set of all non-dominated solution according to Definition 9. Based on Proposition 4 (property 2), we may reformulate this algorithm as the search for subsets of attributes S such that $C^S_{\succ}$ is the identity matrix (i.e., S is a non-dominated solution). The idea is thus to select attributes a such that there exists a couple of observations $(o, o') \in \Theta^2$ satisfying $C[(o, o'), a] = 1$ in the constraints matrix C and $C[(o, o'), a'] = 0$ for any attributes $a' \neq a$.

Data: C: Constraints matrix of size $|\Theta| \times |\mathcal{A}|$.
Result: $\mathcal{S}ol$: Non-dominated solutions set.

$\mathcal{S}ol = \emptyset$
for $i = 1$ *to* $|\Theta|$ **do**
  //Build a subset of solutions ND_i
  $ND_i = \emptyset$
  forall $j \in \mathcal{A}$ *s.t.* $C[\theta_i, j] = 1$ **do**
    forall $S \in \mathcal{S}ol \backslash ND_i$ **do**
      if $j \in S$ **then**
        $ND_i = ND_i \cup \{S\}$
  //Build a subset of solutions ES_i
  $ES_i = \emptyset$
  forall $j \in \mathcal{A}$ *s.t.* $C[\theta_i, j] = 1$ **do**
    forall $S \in \mathcal{S}ol \backslash ND_i$ **do**
      if $\nexists S' \in ND_i$ *s.t.* $S' \subseteq S \cup \{j\}$ **then**
        $ES_i = ES_i \cup \{S \cup \{j\}\}$
  $\mathcal{S}ol = ND_i \cup ES_i$
return $\mathcal{S}ol$;

Algorithm 1. Computation of non-dominated solutions (NDS).

Algorithm 1 builds incrementally the set of non-dominated solutions $\mathcal{S}ol$ with each element of Θ such that $\Theta = \{\theta_1, \theta_2, ..., \theta_{|\Theta|}\}$. At each iteration, the solutions are updated in order to satisfy the constraint corresponding to θ_i. The main idea consists in distinguishing between solutions that already satisfy this constraint (they are put in a set of non dominated solution ND_i) and those that need to be modified in order to satisfy the constraint, (they are put in the set ES_i, an extended set where an attribute is added to a solution). Note that the modification of these latest solutions is performed by adding one attribute but keeping the non-domination property.

6.2 Computation of all Minimal Solutions

In order to compute the set of all minimal solutions S, we introduce a bound B such that $\forall s \in S$, $|s| \leq B$. If no solution satisfying this bound is found we increase the value of B. In this case, the problem of dominance is not addressed since a

[2] Remind that Θ is a set of couples of observations defined in 4.1 for indexing lines of the constraint matrix C.

minimal solution is always non-dominated (see subsection 3.2). However, we must avoid to compute identical solutions.

In [10], the notion of negative attributes was introduced. The representation of a set of solutions $\mathcal{S}ol = \{S_1, S_2, .., S_n\}$ is extended as $\mathcal{S}ol = \{< S_1|NEG_{S_1} >, < S_2|NEG_{S_2} >, ..., < S_n|NEG_{S_n} >\}$, where $NEG_{S_i} \subseteq \{\neg a|a \in \mathcal{A}\}$. The main idea is to never consider an attribute a in a solution S_i if $\neg a \in NEG_{S_i}$.

```
Data: C: Constraints matrix of size |Θ| × |A|.
nbg: number of groups.
Result: Sol: Minimal solutions set.

b = ⌈log2(nbg)⌉; //upper bound
Sol = ∅
do
    for i = 1 to |Θ| do
        //Build a subset of solutions ND_i
        ND_i = ∅
        forall j ∈ A s.t. C[θ_i, j] = 1 do
            forall < S|NEG_S > ∈ Sol\ND_i do
                if j ∈ S then
                    ND_i = ND_i ∪ {< S|NEG_S >}
        //Build a subset of solutions ES_i
        ES_i = ∅
        forall < S|NEG_S > ∈ Sol\ND_i do
            if |S| < b then
                //Progressive construction of a set NEG
                NEG = NEG_S
                forall j ∈ A s.t. C[θ_i, j] = 1 do
                    if ¬j ∉ NEG_S then
                        ES_i = ES_i ∪ < S ∪ {j}|NEG >
                        //Increment the set NEG
                        NEG = NEG ∪ {¬j}
        Sol = ND_i ∪ ES_i
        if Sol = ∅ then
            b = b + 1
            break
while Sol = ∅;
return Sol;
```

Algorithm 2. Computation of minimal solutions (MWNG).

Algorithm 2 uses the same principle as Algorithm 1. At each iteration i, a set ND_i of non-dominated solutions that satisfy the constraint θ_i is built as well as a set ES_i of solutions that must be incremented, if the size of these solutions is lower than the bound. Example 5 shows how the negative attributes avoid getting identical solutions.

Example 5 Given a set of attributes $\mathcal{A} = \{a, b, c, d\}$ and solution $S = \{b, c\}$ that satisfies the $i - 1$ first constraints, let us consider the next three constraint lines $\theta_i, \theta_{i+1}, \theta_{i+2}$ in the constraints matrix C such that:

Constraints	Attributes			
	a	b	c	d
θ_i	1	0	0	1
θ_{i+1}	1	0	0	0
θ_{i+2}	0	0	0	1

The solution $S = \{b, c\}$ does not satisfy the constraint θ_i and we must increment S with a new attribute. Two cases can be considered for generating two new solutions:

- $S_1 = S \cup \{a\}$
- $S_2 = S \cup \{d\}$.

S_1 and S_2 satisfy now θ_i. Note that S_1 satisfies θ_{i+1} and not θ_{i+2} while S_2 satisfies θ_{i+2} but not θ_{i+1}. We need to increment S_2: $S'_2 = S_2 \cup \{a\} = \{a, b, c, d\}$ and S_1: $S'_1 = S_1 \cup \{d\} = \{a, b, c, d\}$ and we get $S'_1 = S'_2$.

Using the concept of negative attributes, we have to create couples of solutions $< S_1|NEG_{S_1} >=< S \cup \{a\}|\emptyset >$ and $< S_2|NEG_{S_2} >=< S \cup \{d\}|\neg a >$in order to satisfy θ_i. Thus, $< S_1|NEG_{S_1} >$ satisfies θ_{i+1}, but $< S_2|NEG_{S_2} >$ does not. S_2 needs to be incremented as $S'_2 = S_2 \cup \{a\}$ but since $\neg a \in NEG_{S_2}$ we must remove S_2.

When we increment S_1 as $S'_1 = S_1 \cup \{d\}$ in $< S_1|NEG_{S_1} >$ in order to satisfy θ_{i+2}, we get a unique $S'_1 = \{a, b, c, d\}$.

Negative attributes aim thus at avoiding redundancy in solutions computation.

7 Comparing LAD and MCP Approaches

In this section, we compare our approach for solving the MCP with the LAD methodology, described in Section 2.

To this aim, we need to restrict instances of MCP to only two groups, since LAD has been originally defined for two groups P and N (positive and negative observations). According to Definition 3, a pdBf is defined by a pair (P, N),$P, N \subseteq \mathbb{B}^n$ and corresponds to a MCP instance $(\Omega, \mathcal{A}, D, G)$ where:

- $\Omega = P \cup N$
- $G : \Omega \mapsto \{P, N\}$, $P = \{o \in \Omega | G(o) = P\}$ and $N = \{o \in \Omega | G(o) = N\}$.

In LAD, the aim is to find a pattern that covers a maximum number of observations of P and such that no observation of N has this pattern. From our MCP point of view our notion of solution is rather different. Given a solution S of a MCP instance $(\Omega, \mathcal{A}, D, G)$ defined as above with only two groups, the variables on S do not in general correspond to a pattern for the observations in P unless $sim_P(D^S) = 1$. In this case the minimal solution of the MCP and the maximal pattern obviously coincide in term of attributes.

Conversely, given $\{p_k | k \in \mathbb{N}\}$ a set of patterns completely covering the group $P = \{o \in \Omega | G(o) = P\}$, then $\bigcup_k Var(p_k)$ is a solution of the corresponding instance of MCP with two groups P and N.

Remind that, a pattern p completely covers a group if all observations in this group are similar on $Var(p)$.

Remark 1 Let us consider the smallest pattern p that completely covers the group P (i.e. $\nexists p'$ s.t. $|Var(p')| < |Var(p)|$ completely covering the group), then $Var(p)$ is a non-dominated solution of the corresponding instance of MCP, but not necessarily of minimal size.

In order to highlight the differences between the two approaches, let us consider the following example.

Example 6

Observations	Groups	Attributes							
		a	b	c	d	e	f	g	h
1	P	0	1	0	1	0	1	0	0
2		0	1	0	1	0	0	1	0
3		0	1	0	1	0	0	0	1
4	N	0	1	0	1	1	0	0	0
5		0	1	0	0	0	0	0	0
6		0	1	1	1	0	0	0	0
7		0	0	0	1	0	0	0	0
8		1	1	0	1	0	0	0	0

$p = \bar{a} \wedge b \wedge \bar{c} \wedge d \wedge \bar{e}$ is a pattern covering the entire group P. However, $\{f, g, h\}$ is a solution and $|\{f, g, h\}| = 3 < |\{a, b, c, d, e\}| = 5$

Definition 20 Given an instance $(\Omega, \mathcal{A}, D, G)$ and a set $A \subset \mathcal{A}$, $\mathcal{O}_A \subset \Omega$ is a set $\{o \in \Omega | \nexists o' \neq o$ s.t. $D^A[o, .] = D^A[o', .]$ and $G(o) = G(o')\}$.

$\mathcal{O}_A$ is the set of all observations where identical lines projected on A have been deleted.

Remark 2 Let S be a solution of an instance of the MCP and $P = \{o \in \mathcal{O}_S | G(o) = P\}$. $|P|$ is therefore the number of patterns p_k composed by attributes of S ($\forall k$ $Var(p_k) = S$) such that $\bigvee_k p_k$ completely covers the group P. However, the patterns p_k are not necessarily of minimum size.

Thus, the canonical writing of solutions is a set of patterns covering the entire group.

Remark 3 Given S a non-dominated solution of an instance of the MCP with two groups (P and N), $P = \{o \in \mathcal{O}_S | G(o) = P\}$ and $\{p_k | k \in \mathbb{N}\}$ the set of patterns covering the group P s.t. $\forall k$ $Var(p_k) \subset S$, we have $\bigcup_k Var(p_k) = S$.

Remark 4 Given S a minimal solution to the MCP, there exists $\{p_k | k \in \mathbb{N}\}$ a set of patterns covering all groups such that $\bigcup_k Var(p_k)$ is minimal, we have $\bigcup_k Var(p_k) = S$

Thus, the Min-MCP determines the smallest set of characters necessary to build patterns covering all groups.

Example 7

Observations	Groups	Attributes		
		a	b	c
1	P	1	1	1
2		1	1	0
3		0	0	0
4	N	0	0	1
5		1	0	1
6		0	1	1

The smallest solution is $S = \{a, b, c\}$. We have 3 patterns $p_1 = a \wedge b \wedge c$, $p_2 = a \wedge b \wedge \bar{c}$ and $p_3 = \bar{a} \wedge \bar{b} \wedge \bar{c}$ completely covering the group. However there exists p_4 and p_5 such that $p_4 = a \wedge b$ and $p_5 = \bar{b} \wedge \bar{c}$, two patterns such that $Var(p_4), Var(p_5) \subset S$ and such that $p_4 \vee p_5$ completely covers the group P. Moreover, we have $Var(p_4) \cup Var(p_5) = S = \{a, b, c\}$.

8 Experimental Setup

The purpose of our experiments is to precisely study the properties of the solutions found for the MCP according to topological properties of the instances and in particular the characteristics of the groups defined in the instances. Since the purpose is to characterize groups of data, we consider classic methods that can be used to group data into cluster using distances. Our purpose is to study the properties of groups with regards to their similarity as well as the properties of the solutions that we compute with regards to different groups functions on similar observations.

8.1 Clustering Approaches

Clustering algorithms are non-supervised machine learning methods that aim at gathering data into clusters according to their similarity. Among the numerous algorithms that can be used for clustering (see [1]), K-means [21,17] is certainly one of the most popular. Given a set Ω of observations the K-means method aims at building k clusters $S_1, \cdots, S_k$ such that $\bigcup_{1 \leq i \leq k} S_i = \Omega$, optimizing the following objective function:

$$min \sum_{i=1}^{k} \sum_{o \in S_i} ||o - \mu_i||^2$$

where μ_i is the centroid (mean value of the points) of S_i.

The purpose of this method is thus to gather similar observations into clusters centered around a centroid. This method tends to maximize the similarity of each cluster. K-means clustering is a NP-hard optimization problem [14]. Here, we use a classic heuristic algorithm [17]. The centroids may be fixed and the clusters are thus built around these given centroids (supervised approach).

The K-medoids method [18] is a clustering method, close to K-means, which gather the observations around a real observation (i.e., the centroids belong to the observation set Ω).

8.2 Ranking Solutions

Using the algorithms described in Section 6, many non dominated solutions can be computed. Therefore, selecting the most suitable solution is a difficult problem since we lack of objective criteria. Nevertheless, since the purpose of solving a MCP is to characterize groups of observations, we may analyze the influence of the selected attributes of a solution on the topology of the group. Using the measures defined in Section 5.3, given an instance $(\Omega, \mathcal{A}, D, G)$, a solution S could be preferred to a solution S' if:

- observations of a given group with regards to S are more similar than with S' (i.e., $sim_{intra}(D^S) > sim_{intra}(D^{S'})$, or
- observations of different groups with regards to S are more different than with S' (i.e. $diff_{inter}(D^S) > diff_{inter}(D^{S'})$.

The first case corresponds to solutions that tend to focus on attributes that have similar values for the observations of a same group. Note that, the intragroup similarity has an impact on the patterns that can be computed from the LAD point of view. Therefore, with regards to this criterion, for an instance $(\Omega, \mathcal{A}, D, G)$ and a set of solution $\mathcal{S}ol$, we denote $\bar{S} \in \mathcal{S}ol$ the solutions such that $\forall S' \in \mathcal{S}ol$, $sim_{intra}(D^{\bar{S}}) \geq sim_{intra}(D^{S'})$, and $\underline{\mathrm{S}} \in \mathcal{S}ol$ the solutions such that $\forall S' \in \mathcal{S}ol, sim_{intra}(D^{\underline{\mathrm{S}}}) \leq sim_{intra}(D^{S'})$.

8.3 Comparing Group Functions

In this section, we are interested in comparing different possible groupings of observations, i.e. comparing two instances $(\Omega, \mathcal{A}, D, G)$ and $(\Omega, \mathcal{A}, D, G')$. A group function G will be close to another function G' if they tend to group observations similarly regardless of the names (e.g., the associated numbers) of the groups. In our experiment we will compare different group functions and we want to define a clear notion of similarity between these functions with regards to the resulting groups.

Unfortunately indicators such as the Jaccard index is not sufficient to evaluate this similarity because they take into account the value of the groups. Hence, it is difficult to use such indexes if there are more than two groups.

Let us consider two instances $(\Omega, \mathcal{A}, D, G)$ and $(\Omega, \mathcal{A}, D, G')$. $\Delta_{G,G'}$ is the contingency matrix for group functions G and G' , i.e., $\Delta_{G,G'}[i, j] = k$ if and only if there are k observations that belongs to groups i according to G and to group j according to G'. Note that the contingency matrix is sensitive to the value of the group (i.e., index of the group). In order to avoid this drawback we have to consider permutations of this contingency matrix.

Let Φ_n be the set of all integer permutation functions on $1 \dots n$ (n being the number of groups). Given $\varphi \in \Phi$, $\Delta^{\varphi}_{G,G'}$ is the contingency matrix whose rows have been rearranged according to φ.

We define our similarity index between G and G' as:

$$\sigma_{G,G'} = \frac{max_{\varphi \in \Phi}\ diag(\Delta^{\varphi}_{G,G'})}{|\Omega|}$$

where $diag(M)$ is the function that returns the sum of the values of the main diagonal of a matrix M.

Note that $argmax_{\varphi \in \Phi}\ diag(\Delta^{\varphi}_{G,G'})$ returns the perfect transformation between G and G'. Note that, for a given contingency matrix of size $|n| \times |n|$, there are $n!$ possible permutations. Therefore, in presence of many groups, computing σ can be very time consuming.

Example 8 Let us consider G, G', G'' three group functions. We represent here these functions by the following vectors, $G = [1,1,1,1,2,2,2,2,3]$, $G' = [2,2,2,2,3,3,3,3,1]$ and $G'' = [1,1,2,2,1,2,3,3,3]$. $G([o])$ (resp. $G'[o]$ and $G''[o]$) represents the group index assigned to observation o. In this example $|\Omega| = 9$.

We have $\Delta_{G,G'} = \begin{pmatrix} 0 & 4 & 0 \\ 0 & 0 & 4 \\ 1 & 0 & 0 \end{pmatrix}$

There exists $\varphi \in \Phi$ such that $\Delta^{\varphi}_{G,G'} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$

$diag(\Delta^{\varphi}_{X,Y}) = 4 + 4 + 1 = 9$ and thus $\sigma_{X,Y} = \frac{9}{9} = 1$.

We have $\Delta_{G,G''} = \begin{pmatrix} 2 & 2 & 0 \\ 1 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$ and $\sigma_{G,G''} = \frac{4}{9}$.

Therefore, we observe that G and G' are similar (see their definition, despite the values of the assigned groups) while G and G'' are more different.

Note that, for an instance $(\Omega, \mathcal{A}, D, G)$, with $G : \Omega \to \{1, \cdots, nb_G\}$ (i.e., nb_G is the number of groups), and another group function $G' : \Omega \to \{1, \cdots, nb_G\}$, the similarity between G and G' is bounded:

$$\sigma_{G,G'} \in [\frac{\lceil |\Omega|/nb_G \rceil}{|\Omega|}, 1]$$

8.4 Description of the Studied Instances

Real Instances

As mentioned in Introduction, the instances used here have been prepared in cooperation with biologists.

Instances *xanthostgen*, *ra_rep1*, *ra_rep2*, *ra_phy*, *ra_phv*, *ra100_phy*, *ra100_phv* and *ralstophy2s* are datasets corresponding to bacterial strains of phytopathogenic bacteria. Each observation is a bacterial strain, and attributes can be housekeeping gene, resistance gene or type III effectors. Groups are pathovar and the objective of the characterization is to determine the delimitations of species more accurately than with a phylogenetic approach (see [9] for more details).

Instance *Leukemia* is a dataset based on patients suffering from leukemia. Observations are mutated human genomes that are suspected to play a role in leukemia. The objective of the characterization is to identify combination of genes in order to predict the risk of relapse of a patient. The human genome is larger than a bacterium, we have then a number attributes of more important in this instance.

These instances are described in the Table 2, which gives their number of observations and attributes and number of groups. These instances are matrices built from biological data [3].

Table 2 Biological instances

Instances	Observations×Attributes	# Groups
xanthostgen	173×107	33
ra_rep1	112×155	7
ra_rep2	112×155	7
ra_phy	113×155	4
ra_phv	109×155	8
ra100_phy	113×99	4
ra100_phv	109×99	8
ralstophy2s	112×34	4
Leukemia	35×8389	2

Handmade instances

In order to consider particular cases with specific extremal properties, we build two instances. Let us consider $|\Omega| = 34$ and $|\mathcal{A}| = 17$ ($\mathcal{A} = \{a, b, c, d, ..., p, q\}$).

$$\begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

We consider two group functions $G_1, G_2 : \Omega \mapsto \{1, 0\}$ such that:

- $G_1(i) = 1$ if $i < 18$, $G_1(i) = 2$ if $i \geq 18$.
- $G_2(i) = 1$ if i is odd, $G_2(i) = 2$ if i is even.

We have the following values:

G_g	Minimal sol size	sim_{over}	intragroup similarity	intergroup difference
G_1	3	0.485	0.882	0.889
G_2	8	0.485	0.478	0.508

Note that the patterns $p_1 = a \wedge b$ and $p_2 = c \wedge d$ cover together the group 1 with the group function G_1.
The patterns $p_3 = b \wedge d \wedge f \wedge h \wedge j \wedge l \wedge n \wedge p$ and $p_4 = \bar{a} \wedge \bar{c} \wedge \bar{e} \wedge \bar{g} \wedge \bar{i} \wedge \bar{k} \wedge \bar{m} \wedge \bar{o} \wedge \bar{q}$ cover together the group 1 with the group fuction G_2.

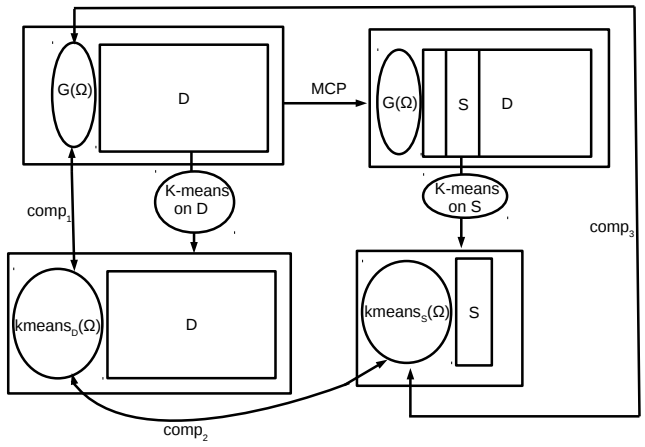
Here we can see that for the same data matrix, a group of observations giving a high intragroup similarity and high intergroup difference will give us smaller solutions.

[3] http://www.info.univ-angers.fr/~gh/Idas/Ccd/ce_f.php

8.5 Experimental Protocol

Figure 1 presents our experimental protocol in order to precisely examine the notion of solution in the context of MCP. In particular we are interested in studying if the attributes identified in solutions as well as the size of the solutions can be related to structural properties of the instance.

Fig. 1 Experimental protocol



For each instance $(\Omega, \mathcal{A}, D, G)$ we consider several group functions:

- the original group function G corresponding to groups defined by experts in real cases (G on top left and top right in Figure 1);
- group functions obtained by clustering methods on initial data ($kmeans_D$ on bottom left in Figure 1) and on data reduced to attributes in solutions ($kmeans_S$ on bottom right in Figure 1). The following variants are considered:
  - group function corresponding to groups generated by a K-means method (see Section 8.1);
  - group function corresponding to groups generated by a K-means with fixed centroids, called here center k-means;
  - group function corresponding to groups generated by a K-medoid method (see Section 8.1);
  - randomly generated group function.

Note that all these functions consider the same number of groups, which corresponds to the original group function provided in the initial instances.

The purpose of this first analysis, corresponding to $comp_1$ in Figure 1, is to assess the impact of the group function on the size of solutions as well as on the similarity of the groups.

For each instance with the above-mentioned group functions, we compute solutions (MCP arrow in Figure 1). From these solutions, we consider again the same groups but restricted to the set of attributes that appear in the solutions.

The comparisons $comp_2$ and $comp_3$ correspond therefore to the analysis of the impact of the solutions attributes on the groups issued from the different group functions with regards to the original group function ($comp_3$) and with the group functions resulting from other classification methods ($comp_2$).

9 Experimental Results

The following subsections are directly related to the different steps of our experimental protocol (represented by arrows on Figure 1). Therefore, it will be recalled when relevant.

9.1 Computation of Minimal Solutions

Experimetal protocol : it corresponds to the MCP arrow on Figure 1 but the computation of MCP solutions is also performed on groups obtained by the k-means algorithm (denoted $kmeans_D(\Omega)$).

Table 3 provides the size of the minimal solutions for each instance with different group functions.

We observe that randomly generated groups are more difficult to characterize and require thus more attributes, since as expected no regularity can be exploited from the data. The other results show that the original groups and the groups obtained by the classification algorithm are not so different in terms of MCP solution sizes. Therefore, further analysis must be performed on these groups.

9.2 Comparison of Group Functions

Experimental protocol : it corresponds to the $comp_1$ arrow on Figure 1.

In order to compare the different group functions presented in Section 8.5, we use the objective function of k-means method (see Section. 8.1). Note that the value k is set to the number of groups given by the original group function. The purpose is to evaluate the coherence of the groups with regards to classic classification criteria. Given an instance, for each group, we compute the value of the objective function of the k-means method. This value provides us informations about the structure of the groups in terms of concentration of observations. Note that, the higher the size of the data matrix is (i.e., in terms of number of observations and attributes), the higher the value of the objective function is. We normalize these values according to the number of observations and the number of attributes. These values are reported in Table 4.

Table 3 Size of minimal solutions

Instances	Group Function	Minimal solution size
xanthostgen	Original	13
xanthostgen	K-medoid	11
xanthostgen	K-means	12
xanthostgen	center k-means	12
xanthostgen	Random	33
ra_rep1	Original	12
ra_rep1	K-medoid	9
ra_rep1	K-means	7
ra_rep1	center k-means	7
ra_rep1	Random	16
ra_rep2	Original	12
ra_rep2	K-medoid	9
ra_rep2	K-means	8
ra_rep2	center k-means	8
ra_rep2	Random	16
ra_phy	Original	6
ra_phy	K-medoid	8
ra_phy	K-means	6
ra_phy	center k-means	6
ra_phy	Random	15
ra_phv	Original	6
ra_phv	K-medoid	9
ra_phv	K-means	8
ra_phv	center k-means	7
ra_phv	Random	16
ra100_phy	Original	7
ra100_phy	K-medoid	8
ra100_phy	K-means	7
ra100_phy	center k-means	7
ra100_phy	Random	16
ra100_phv	Original	9
ra100_phv	K-medoid	10
ra100_phv	K-means	9
ra100_phv	center k-means	10
ra100_phv	Random	16
ralstophy2s	Original	8
ralstophy2s	K-medoid	7
ralstophy2s	K-means	5
ralstophy2s	center k-means	3
ralstophy2s	Random	13
Leukemia	Original	2
Leukemia	K-mdoid	4
Leukemia	K-means	3
Leukemia	center k-means	3
Leukemia	Random	2

Table 5 provides the similarity between group functions with the original group function according to the similarity index defined in Section 8.3 ($comp_1$ in Fig 1). Note that for *xanthostgen*, the number of groups is too high to compute the index. Of course the similarity is 1 for the original group functions.

As expected, the random group functions are very different from the original group functions. We may observe that group functions obtained by classification are indeed rather different from the original group functions, while we have ob-

Table 4 Value of k-means objective function.

Instances	Group functions	Minimal solution size	Initial k-means objective function	Normalized k-means objective function
xanthostgen	Original	13	186.500	0.019
xanthostgen	K-medoid	11	136.768	0.014
xanthostgen	K-means	12	155.385	0.016
xanthostgen	center k-means	12	154.393	0.016
xanthostgen	Random	33	990.117	0.103
ra_rep1	Original	12	755.022	0.092
ra_rep1	K-medoid	9	583.971	0.071
ra_rep1	K-means	7	570.379	0.070
ra_rep1	center k-means	7	570.379	0.070
ra_rep1	Random	16	982.077	0.120
ra_rep2	Original	12	754.501	0.092
ra_rep2	K-medoid	9	583.971	0.071
ra_rep2	K-means	8	590.362	0.072
ra_rep2	center k-means	8	575.717	0.070
ra_rep2	Random	16	981.837	0.120
ra_phy	Original	6	795.025	0.097
ra_phy	K-medoid	8	694.224	0.085
ra_phy	K-means	6	665.324	0.081
ra_phy	center k-means	6	665.324	0.081
ra_phy	Random	15	1003.030	0.123
ra_phv	Original	6	542.583	0.072
ra_phv	K-medoid	9	487.656	0.065
ra_phv	K-means	8	500.239	0.066
ra_phv	center k-means	7	480.067	0.064
ra_phv	Random	16	892.476	0.118
ra100_phy	Original	7	516.776	0.097
ra100_phy	K-medoid	8	416.205	0.078
ra100_phy	K-means	7	405.415	0.076
ra100_phy	cent kmeans	7	418.636	0.078
ra100_phy	Random	16	594.883	0.111
ra100_phv	Original	9	373.076	0.074
ra100_phv	K-medoid	10	300.570	0.060
ra100_phv	K-means	9	304.891	0.060
ra100_phv	cent kmeans	10	295.815	0.059
ra100_phv	Random	16	546.836	0.108
ralstophy2s	Original	8	205.019	0.122
ralstophy2s	K-medoid	7	169.080	0.101
ralstophy2s	K-means	5	162.639	0.097
ralstophy2s	cent kmeans	3	162.482	0.097
ralstophy2s	Random	13	247.306	0.147
Leukemia	Original	2	8798.400	0.144
Leukemia	K-mdoid	4	8756.820	0.143
Leukemia	K-means	3	8734.740	0.143
Leukemia	cent kmeans	3	8717.610	0.143
Leukemia	Random	2	8813.620	0.144

served previously that these group functions generate groups whose cohesions are rather similar (according to the k-mean objective function).

In order to further investigate the characteristics of the groups induced by the group functions, we evaluate instances with more topological criteria presented in Section 5.3.

Table 5 Similarity $\sigma_{G,O}$ between the different group methods and the initial group function (O).

Instance	Group Function G	Minimal solution size	similarity $\sigma_{G,O}$ with original group function
ra_rep1	Original	12	1
ra_rep1	K-medoid	9	0.437
ra_rep1	K-means	7	0.536
ra_rep1	center k-means	7	0.536
ra_rep1	Random	16	0.232
ra_rep2	Original	12	1
ra_rep2	K-medoid	9	0.429
ra_rep2	K-means	8	0.420
ra_rep2	center k-means	8	0.5
ra_rep2	Random	16	0.214
ra_phy	Original	6	1
ra_phy	K-medoid	8	0.589
ra_phy	K-means	6	0.562
ra_phy	center k-means	6	0.562
ra_phy	Random	15	0.321
ra_phv	Original	6	1
ra_phv	K-medoid	9	0.667
ra_phv	K-means	8	0.722
ra_phv	center k-means	7	0.796
ra_phv	Random	16	0.259
ra100_phy	Original	7	1
ra100_phy	K-medoid	8	0.457
ra100_phy	K-means	7	0.514
ra100_phy	center kmeans	7	0.619
ra100_phy	Random	16	0.505
ra100_phv	Original	9	1
ra100_phv	K-medoid	10	0.584
ra100_phv	K-means	9	0.505
ra100_phv	center kmeans	10	0.594
ra100_phv	Random	16	0.396
ralstophy2s	Original	8	1
ralstophy2s	K-medoid	7	0.575
ralstophy2s	K-means	5	0.697
ralstophy2s	center kmeans	3	0.644
ralstophy2s	Random	13	0.370
Leukemia	Original	2	1
Leukemia	K-medoid	4	0.514
Leukemia	K-means	3	0.543
Leukemia	center kmeans	3	0.571
Leukemia	Random	2	0.771

9.3 Deeper Analysis of Group Functions

Table 6 sumarizes characteristics for the instances with the different groups functions as previously defined. We consider the instances xanthostgen, ra100_phy, ra100_phv, ralstophy2s and Leukemia. For each group function, the following characteristics are recorded : minimal solution size, overall similarity of the observations, intragroup similarity and intergroup difference (see definitions in Section 5.3).

Table 6 Similarities (as defined in Section 5.3) for the different group functions

Instances	Group functions	minimal solution size	sim_{over}	sim_{intra}	$diff_{inter}$
Ex.	G_1	3	0.485	0.882	0.889
Ex.	G_2	8	0.485	0.478	0.508
xanthost	Original	13	0.719	0.942	0.289
xanthost	K-med	11	0.719	0.960	0.289
xanthost	K-means	12	0.719	0.954	0.289
xanthost	center K-means	12	0.719	0.954	0.288
xanthost	Random	33	0.719	0.708	0.281
ra_rep1	Original	12	0.747	0.793	0.270
ra_rep1	K-med	9	0.747	0.855	0.273
ra_rep1	K-means	7	0.747	0.870	0.280
ra_rep1	center K-means	7	0.747	0.870	0.280
ra_rep1	Random	16	0.747	0.742	0.253
ra_rep2	Original	12	0.747	0.792	0.269
ra_rep2	K-med	9	0.747	0.855	0.273
ra_rep2	K-means	8	0.747	0.857	0.277
ra_rep2	center K-means	8	0.747	0.865	0.277
ra_rep2	Random	16	0.747	0.741	0.252
ra_phy	Original	6	0.747	0.808	0.303
ra_phy	K-med	8	0.747	0.830	0.285
ra_phy	K-means	6	0.747	0.837	0.284
ra_phy	center K-means	6	0.747	0.837	0.284
ra_phy	Random	15	0.747	0.747	0.253
ra_phv	Original	6	0.748	0.847	0.293
ra_phv	K-med	9	0.748	0.867	0.273
ra_phv	K-means	8	0.748	0.871	0.281
ra_phv	center K-means	7	0.748	0.878	0.280
ra_phv	Random	16	0.748	0.742	0.251
ra100phy	Original	7	0.757	0.808	0.303
ra100phy	K-med	8	0.757	0.839	0.270
ra100phy	K-means	7	0.757	0.843	0.271
ra100phy	center K-means	7	0.757	0.838	0.278
ra100phy	Random	16	0.757	0.774	0.255
ra100phv	Original	9	0.757	0.833	0.271
ra100phv	K-med	10	0.757	0.872	0.263
ra100phv	K-means	9	0.757	0.871	0.262
ra100phv	center K-means	10	0.757	0.878	0.264
ra100phv	Random	16	0.757	0.764	0.246
ralsto	Original	8	0.676	0.854	0.253
ralsto	K-med	7	0.676	0.790	0.363
ralsto	K-means	5	0.676	0.797	0.370
ralsto	center K-means	3	0.676	0.795	0.366
ralsto	Random	13	0.676	0.678	0.326
Leuk	Original	2	0.694	0.695	0.309
Leuk	K-med	4	0.694	0.696	0.308
Leuk	K-means	3	0.694	0.697	0.309
Leuk	center K-means	3	0.694	0.697	0.309
Leuk	Random	2	0.694	0.693	0.304

Note that the overall similarity is computed on the whole set of observations and is of course identical for all different group functions. We first remark that this overall similarity is rather high, which may be explained by the fact that the observations are not randomly generated and share similar values on a sufficient number of attributes. This assumption is rather reasonable in particular when attributes are genes, whose presence/absence is indeed often similar for different individuals.

Note that the intragroup similarity of random group functions is close to the overall similarity. This is coherent since random group functions gather initial observations uniformly, keeping thus their initial similarity.

The intragroup similarity is high for initial group functions as well as for k-means based group functions, which corroborates the previous analysis. In fact, these group functions have similar topological properties in terms of group similarity and solution length. Nevertheless the resulting groups are different.

Concerning the intergroup difference, the scores are rather low. This can be explained by the high global similarity of the observations. It is indeed difficult to generate very different groups of observations.

Note that the *Leukemia* instance is slightly different. For each group function, even for the random group function, the intragroup similarity and intergroup difference are much closer. We may observe that minimal solutions sizes are very small, especially with respect to the set of initial attributes which is very large. We will discuss these results later.

We have observed that original group functions have topological properties close to groups generated by k-means algorithms. Nevertheless, the resulting groups are very different in terms of observations. Therefore, we may conclude that the initial groups cannot be obtained using k-means algorithms with a classic Hamming distance metric. We could try to modify this distance but we have a priori no available information that can be used to select the most suitable attributes.

A remaining question is thus: Does our attribute reduction approach allow us to characterize groups that could be obtained by clustering algorithms if the suitable attributes had been identified *a priori* ?

9.4 Analysis of Solutions

Experimental protocol : it corresponds to the $comp_2$ and $comp_3$ arrows on Figure 1. $comp_2$ corresponds to the comparison between groups obtained by clustering on the initial observations and groups obtained by clustering on the observations reduced to the attributes of optimal solutions. $comp_3$ corresponds to the comparison of the groups obtained by clustering on the observations reduced to the attributes of optimal solutions and the original groups.

In this part, we focus on the solutions computed for the different instances, using different group functions. According to our experimental protocol presented in Figure 1, we are interested in analyzing if attributes of solutions may be exploited by k-means algorithms in terms of similarity.

To this aim, we apply the clustering algorithm on the observations restricted to the attributes that appear in solutions (obviously the distance values change). Then, we compare the similarity of the resulting groups obtained by clustering with the original group functions.

Before considering these comparisons, we have to analyze solutions. Given instance $(\Omega, \mathcal{A}, D, G)$, solving the minimum multiple characterization problem lead to many optimal solutions. For each solution, the resulting intragroup similarity can be different.

Table 7 shows intragroup similarity, overall similarity and intergroup difference for the solution $\bar{S}$ that maximizes the intragroup similarity and for the solution

$\underline{S}$ that minimizes this intragroup similarity. The last column corresponds to the solution S' that provides the highest similarity of a group, i.e. $S' = argmax_{S \in Sol}(max_g sim_g(D^{S'}))$ (see Definition 15).

Table 7 Solutions similarity

Instance	Minimal sol size	$sim_{intra}(D^{\bar{S}})$ max	$sim_{over}(D^{\bar{S}})$	$diff_{inter}(D^{\bar{S}})$	$sim_{intra}(D^{\underline{S}})$ min	$sim_{over}(D^{\underline{S}})$	$diff_{inter}(D^{\underline{S}})$	$sim_g(D^{S'})$ max
Example G_1	3	0.882	0.485	0.889	0.882	0.485	0.889	0.882
Example G_2	8	0.485	0.485	0.516	0.485	0.485	0.516	0.485
xanthostgen	13	0.958	0.664	0.346	0.908	0.646	0.363	1
ra_rep1	12	0.671	0.603	0.422	0.636	0.595	0.420	0.75
ra_rep2	12	0.691	0.637	0.382	0.605	0.561	0.454	0.917
ra_phy	12	0.869	0.725	0.392	0.843	0.594	0.607	0.906
ra_phv	12	0.883	0.630	0.474	0.815	0.565	0.538	0.921
ra100_phy	7	0.841	0.710	0.388	0.701	0.576	0.516	0.875
ra100_phv	9	0.782	0.620	0.439	0.670	0.549	0.494	1
ralstophy2s	8	0.736	0.638	0.422	0.650	0.562	0.493	0.770
Leukemia	2	0.992	0.889	0.5	0.537	0.503	0.629	1

Obviously, we have $sim_{intra}(D^{\bar{S}}) \geq sim_{intra}(D^{\underline{S}})$. In all cases, except for *ra_rep1*, we observe that $diff_{inter}(D^{\underline{S}}) \geq diff_{inter}(D^{\bar{S}})$. Moreover, the overall similarity value is higher for $D^{\bar{S}}$ than for $D^{\underline{S}}$.

Note that when the similarity of a group is high, it means that we may expect that the group can be covered by a few patterns (see Section 7).

The k-means algorithm used here is based on the Euclidean distance between observations. But we have seen that the initial group functions cannot be explained by distance-based similarity. We now want to investigate if, once projected on the attributes selected in the solutions obtained for the MCP, the resulting groups of observations exhibit different characteristics with regards to clustering approaches. In other words, we want to check if groups obtained by clustering on the reduced set of attributes are similar to initial groups.

Table 8 presents the similarity σ between group functions. The smaller the minimal solution size is, the larger the similarity between the original group function and the k-means based group functions computed on solutions is. The similarity between k-means based group functions and the original group function is close to the similarity between initial k-means based group functions and k-means group functions on solutions.

9.5 Experimental Conclusions

We summarize now our main observations :

– The similarity σ between the original groups and those obtained by the k-means method presented in Section 8.1 ($comp_1$ and $comp_3$ in Fig 1) is very

Table 8 Comparison of group functions

Instances	Group functions method	Similarity σ with original	Similarity σ with k-means on sol	Similarity σ with center k-means on sol
Example	G_1	1	1	1
Example	K-means	1	1	1
Example	center k-means	1	1	1
ra_rep1	Original	1	0.339	0.455
ra_rep1	K-means	0.536	0.464	0.464
ra_rep1	center k-means	0.536	0.464	0.464
ra_rep2	Original	1	0.375	0.411
ra_rep2	K-means	0.420	0.491	0.455
ra_rep2	center k-means	0.5	0.562	0.509
ra_phy	Original	1	0.652	0.875
ra_phy	K-means	0.562	0.491	0.634
ra_phy	center k-means	0.562	0.491	0.634
ra_phv	Original	1	0.787	0.889
ra_phv	K-means	0.722	0.694	0.713
ra_phv	center k-means	0.796	0.778	0.843
ra100_phy	Original	1	0.743	0.771
ra100_phy	K-means	0.514	0.581	0.619
ra100_phy	center k-means	0.619	0.667	0.762
ra100_phv	Original	1	0.545	0.663
ra100_phv	K-means	0.505	0.475	0.564
ra100_phv	center k-means	0.594	0.465	0.594
ralstophy2s	Original	1	0.534	0.644
ralstophy2s	K-means	0.699	0.644	0.740
ralstophy2s	center k-means	0.644	0.699	0.808
Leukemia	Original	1	0.943	1
Leukemia	K-means	0.543	0.543	0.543
Leukemia	center k-means	0.571	0.514	0.571

low (despite close rates of intra-group similarity and very close rates of inter-group difference).
- The same observation is valid if k-means is applied on the initial set of attributes ($comp_1$ in Fig 1) or only on attributes used in the best solutions ($comp_3$ in Fig 1).
- Except for the *Leukemia* instance, using the attribute of the minimal MCP solutions does not allow to find the original groups with k-means ($comp_3$ in Fig 1).
- Classifications on attributes that belong to solutions remain different from initial groups ($comp_2$ in Fig 1).

Logical data characterization provides thus an alternative knowledge on groups. This methodology rather provides a combinatorial identification of groups signatures that may reveal alternative relationships between attributes than classic models based on statistical occurrences of values.

10 Conclusion

In this paper we have presented a complete methodology for the characterization of multiple groups of Boolean observations using minimal set of attributes (i.e., Boolean variables). This methodology allows the user to identify combinations of attributes that cannot be easily extracted or observed with regards to topological

measures. In particular, we have observed on different sets of real biological data that groups of observations identified by experts cannot be simply obtained nor explained using classic clustering techniques. While these instances have grouping properties that are closed to groups resulting from clustering algorithms, more complex relationships between attributes can be identified between observations. Instead of computing common patterns that are present in observations, our approach rather consists in identifying a characterization formula that can be used to explain the properties of a group of observations or to predict the membership of new incoming observations to existing groups.

We have proposed algorithms that allows the expert to compute complete sets of possible solutions and thus that could help him in better identifying and understanding hidden relationships within data.

References

1. Aggarwal, C. C., Reddy, C. K., 2013. Data clustering: algorithms and applications. CRC Press.
2. Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Acm sigmod record. Vol. 22. ACM, pp. 207–216.
3. Alexe, G., Alexe, S., Axelrod, D., Hammer, P. L., Weissmann, D., 2005. Logical analysis of diffuse large b-cell lymphomas. Artificial Intelligence in Medicine 34 (3), 235 – 267.
4. Alexe, G., Alexe, S., Axelrod, D. E., Bonates, T. O., Lozina, I. I., Reiss, M., Hammer, P. L., 2006. Breast cancer prognosis by combinatorial analysis of gene expression data. Breast Cancer Research 8 (4), 1–20.
5. Alexe, G., Alexe, S., Bonates, T. O., Kogan, A., 2007. Logical analysis of data - the vision of Peter L. Hammer. Annals of Mathematics and Artificial Intelligence 49 (1-4), 265–312.
6. Bennane, A., Yacout, S., 2012. Lad-cbm; new data processing tool for diagnosis and prognosis in condition-based maintenance. Journal of Intelligent Manufacturing 23 (2), 265–275.
7. Boros, E., Crama, Y., Hammer, P. L., Ibaraki, T., Kogan, A., Makino, K., 2011. Logical analysis of data: classification with justification. Annals of Operations Research 188 (1), 33–61.
8. Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., 1997. Logical analysis of numerical data. Mathematical Programming 79, 163–190.
9. Boureau, T., Kerkoud, M., Chhel, F., Hunault, G., Darrasse, A., Brin, C., Durand, K., Hajri, A., Poussier, S., Manceau, C., Lardeux, F., Saubion, F., Jacques, M.-A., 2013. A multiplex-pcr assay for identification of the quarantine plant pathogen xanthomonas axonopodis pv. phaseoli. Journal of Microbiological Methods 92 (1), 42 – 50.
10. Chambon, A., Boureau, T., Lardeux, F., Saubion, F., Le Saux, M., 2015. Characterization of multiple groups of data. In: Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on. IEEE, pp. 1021–1028.
11. Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering 40 (1), 16 – 28, 40th-year commemorative issue.
12. Chhel, F., Lardeux, F., Saubion, F., Zanuttini, B., 2013. Application du problème de caractérisation multiple à la conception de tests de diagnostic pour la biologie végétale. Revue d'Intelligence Artificielle 27 (4-5), 649–668.
13. Chikalov, I., Lozin, V., Lozina, I., Moshkov, M., Nguyen, H., Skowron, A., Zielosko, B., 2013. Logical analysis of data: Theory, methodology and applications. In: Three Approaches to Data Analysis. Vol. 41 of Intelligent Systems Reference Library. Springer Berlin Heidelberg, pp. 147–192.
14. Dasgupta, S., 2008. The hardness of k-means clustering. Department of Computer Science and Engineering, University of California, San Diego.
15. Dupuis, C., Gamache, M., Jean-François, P., 2012. Logical analysis of data for estimating passenger show rates at air canada. Journal of Air Transport Management 18 (1), 78–81.
16. Hammer, P. L., Bonates, T. O., 2006. Logical analysis of data - an overview: From combinatorial optimization to medical applications. Annals of Operations Research 148 (1), 203–225.

17. Hartigan, J. A., Wong, M. A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1), 100–108.
18. Kaufman, L., Rousseeuw, P. J., 1990. Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, Wiley, 68–125.
19. Kholodovych, V., Smith, J. R., Knight, D., Abramson, S., Kohn, J., Welsh, W. J., 2004. Accurate predictions of cellular response using qspr: a feasibility test of rational design of polymeric biomaterials. Polymer 45 (22), 7367 – 7379.
20. Kumar, V., Abbas, A. K., Fausto, N., Aster, J. C., 2014. Robbins and Cotran pathologic basis of disease. Elsevier Health Sciences.
21. MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. Oakland, CA, USA., pp. 281–297.
22. Makino, K., Hatanaka, K., Ibaraki, T., 1999. Horn extensions of a partially defined boolean function. SIAM Journal on Computing 28 (6), 2168–2186.