Math. & Sci. hum. / Mathematics and Social Sciences (43e année, n° 169, *2005*(1), p. 77-104)

STRUCTURING PROBABILISTIC DATA BY GALOIS LATTICES

Paola BRITO[1], Géraldine POLAILLON[2]

SUMMARY – *In this paper we address the problem of organising probabilistic data by Galois concept lattices. Two lattices are proposed, the union lattice and the intersection lattice, corresponding to two distinct semantics, by choosing accordingly the join and meet operators. A new algorithm is proposed to construct the concept lattice. Two real data examples illustrate the presented approach.*

KEYWORDS : Galois lattice, Probabilistic data, Conceptual clustering

RÉSUMÉ – Organisation de données probabilistes par des treillis de Galois

Dans cet article, nous nous intéressons à l'organisation de données probabilistes par des treillis de Galois. Deux correspondances de Galois sont établies, en définissant de façon appropriée les opérateurs de généralisation et d'extension. Ces correspondances permettent de construire deux treillis, appelés treillis de l'union et treillis de l'intersection, correspondant à ces deux correspondances de Galois. Un nouvel algorithme de construction de treillis est proposé. Deux exemples sur des données réelles illustrent l'approche présentée.

MOTS-CLÉS – Treillis de Galois, Données probabilistes, Classification conceptuelle

1. INTRODUCTION

After the definition, by Birkhoff, of Galois connections [Birkhoff, 1940], the Galois lattice associated to an object/variable correspondence has quickly appeared as an important tool in binary data analysis. Barbut and Monjardet [1970] emphasise the interest of Galois connections in the study of a correspondence. Since the eighties, the importance of the Galois lattice of a relation has started to be widely recognised. In fact, many theoretical and algorithmic developments have been accomplished, on the one hand by the group of Wille on Formal Concept Analysis [Wille, 1982; Ganter, Wille, 1999], and on the other hand by Duquenne [Duquenne, Guigues, 1986]. These studies have used the lattice theory for the analysis, organisation and interpretation of data. Galois lattices may be considered as a clustering, allowing the structured clusters to be identified and automatically interpreted. These lattices also reveal the links between the objects, the variables, and between objects and variables. The Galois lattice has a formally defined structure that does not depend on external parameters, on the ordering of instances or algorithmic details. A great deal of research work has been accomplished recently on Galois lattices: on the reduction of the lattice by pruning [Godin *et al*, 1995], [Guénoche, 1993; Mephu Nguifo, 1993],

[1] Universidade do Porto, Faculdade de economia, rua Dr. Roberto Frias, 4200-464 Porto, Portugal, mpbrito@fep.up.pt

[2] SUPELEC – service informatique, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, geraldine.polaillon@supelec.fr

knowledge acquisition [Wille, 1990], rule generation [Duquenne, Guigues, 1986; Duquenne, 1987], frequent items sets [Agrawal *et al*, 1993], [Pasquier *et al*, 1999; Stumme *et al*, 2001], and the extension to richer representations such as conceptual graphs [Bournaud, 1996], imprecise and structured data [Girard, Ralambondrainy, 1999], and fuzzy sets [Burusco, Fuentes-Gonzales, 1998; Herrmann, Hölldobler, Strohmaier, 1996; Pollandt, 1997; Wolff, 2002].

The need to consider data that contain information which cannot be represented in a classical data matrix, together with the objective of designing methods that produce results interpretable in terms of the input variables, lead to the development of Symbolic Data Analysis. Symbolic data extend the classical tabular model, where each individual takes exactly one value for each variable, by allowing multiple, possibly weighted, values for each variable. New variable types have been introduced, which allow us to represent variability and/or uncertainty present in the data: multi-valued variables, interval variables and modal variables [Bock, Diday, 2000]. A variable is called set-valued if its "values" are nonempty subsets of the underlying domain; it is multi-valued if its values are finite subsets of the domain and it is an interval variable if its values are intervals of a linear order (for instance the real numbers). A modal variable is a multi-state variable where we are given a category set for each element and a frequency or probability for each category, which indicates how frequent or likely that category is for this element. In the case where an empirical distribution is given, the variable is called a histogram variable. Data described by modal variables are called probabilistic data. This kind of data often arises in many practical applications, for instance, when it is wished to express uncertainty, or on summarising data from a survey. The main objective of Symbolic Data Analysis is to extend data analysis techniques to such symbolic data, in such a way that both input data and output results may be expressed within the same formalism, based on the notion of 'symbolic object'. Symbolic Data Analysis underwent great improvement with the European projects "Symbolic Official Data Analysis System (SODAS)" and "Analysis System of Symbolic Official data (ASSO)"; as the result of these projects a software package *SODAS* has been developed, [Bock, Diday, 2000].

The problem of extending Galois connections and Galois lattices to symbolic data was first addressed by Brito [1991, 1994] and further developed by Polaillon [1998(a), 1998(b); Polaillon, Diday, 1999]. In this paper, our aim is to define the tools that will allow constructing Galois lattices directly on probabilistic data, without any prior transformation. To do so, we use the framework of Symbolic Data Analysis, which allows representing within the same formalism the input data and the obtained concepts.

In Section 2, we start by explaining what is meant by probabilistic data, and how it is formalized in the context of Symbolic Data Analysis. In Section 3 we define two Galois connections on a set of probabilistic objects, and the corresponding concept lattices, and extend the notion of "complete symbolic object" to this data. In Section 4, we present a new algorithm for the construction of the concept lattice. Comparisons with results published by other authors and two applications, made on real data, are presented and discussed in Section 5.

2. PROBABILISTIC DATA

In classical data analysis, data are represented in a rectangular matrix, where n individuals are represented in rows and p variables in columns and each individual takes exactly one value for each variable. However, it is often the case that this model is too simple to

represent real data, which appear to be more complex. We consider the case where variables are discrete and individuals present a distribution rather than a single value for each variable. This may be a probability distribution, when data are uncertain, or a frequency distribution when data result from aggregation.

As an example consider the case of describing towns for which the variable weather is uncertain, and takes its different values with some probability; then in a town T the weather will be sunny with 60 % probability, cloudy with 30 % probability or rainy with 10 % probability.

Another example arises when we want to summarise data issued from a survey, and individual descriptions have to be aggregated. For instance, a region may be described by saying that 50 % of people are working, 20 % are students, 10 % are retired and the remaining 20 % are in another situation.

Variables for which a distribution is given have been called "Modal variables" [Boc, Diday, 2000]. A modal variable Y with finite domain $O = \{y_1, \ldots, y_k\}$ on a set $E = \{\omega_1, \omega_2, \ldots\}$, where each y_ℓ is called a *category*, is a multi-state variable where, for each element ω of E, we are given a category set $Y(\omega) \subseteq O$ and, for each $y \in Y(\omega)$ a frequency $f(y)$ or probability $p(y)$ which indicates how frequent or likely that category is for this element ω. In the previous example, the domain is the set of the possible professional status, and "student" is one category. Formally,

DEFINITION 1 [Bock, Diday, 2000]

A **modal variable** Y on a set $E = \{\omega_1, \omega_2, \ldots\}$ with domain O is a mapping $Y(\omega) = (U(\omega), \pi_\omega)$, for $\omega \in E$, where π_ω is a measure (frequency, probability or weight) distribution on the domain O of possible observation values (completed by a suitable σ-field), and $U(\omega) \subseteq O$ is the support of π_ω in the domain O.

Generally, the support $U(\omega)$ can be omitted from the definition and so a modal variable can be seen as a mapping $Y : E \to M(O)$, from E into the family $M(O)$ of all non-negative measures π on O, with values $Y(\omega) = \pi_\omega$.

EXAMPLE 1. Let $e = \{\text{region}_1, \text{region}_2, \ldots, \text{region}_{10}\}$ be a set of ten regions for which the gender distribution is known. let y = "gender" with domain $o = \{\text{male, female}\}$, such that $y: e \to m(o)$, $y(\text{region}_i) = \{\text{male } (p_1(i)), \text{female } (p_2(i))\}$ is the gender frequency distribution in region_i, i=1,…, 10. y is a modal variable. we may have, for instance, $y(\text{region}_1) = \{\text{male (0.6), female (0.4)}\}$.

Such kind of data cannot be considered within the classical paradigm of rectangular matrices. A suitable formalism must be adopted to represent and analyse the data. Symbolic Data Analysis provides this framework.

DEFINITION 2

An **modal event** is a statement of the form

$$e = [Y(\omega)\ R\ \{y_1(p_1), y_2(p_2), \ldots, y_k(p_k)\}]$$

where $O = \{y_1, y_2, \ldots, y_k\}$ is the domain of Y, and p_j is the probability, frequency or weight of y_j. It is not imposed that $p_1 + p_2 + \ldots + p_k = 1$. R is a relation on the set of distributions on O. We shall consider the following relations:

1. "~" such that $[Y(\omega) \sim \{y_1(p_1),\ldots, y_k(p_k)\}]$ is true iff $p_\ell(\omega) = p_\ell,\ \ell = 1,\cdots,k$;
2. "$\subseteq$"such that $[Y(\omega) \subseteq \{y_1(p_1),\ldots, y_k(p_k)\}]$ is true iff $p_\ell(\omega) \leq p_\ell,\ \ell = 1,\cdots,k$;
3. "$\geq$"such that $[Y(\omega) \geq \{y_1(p_1),\ldots, y_k(p_k)\}]$ is true iff $p_\ell(\omega) \geq p_\ell,\ \ell = 1,\cdots,k$.

A **modal object** is a conjunction of modal events.

EXAMPLE 2. let again e ={region$_1$, region$_2$, …, region$_{10}$} be a set of regions for which the gender and education level distributions are known. suppose we wish to describe a region where the gender distribution is uniform and where 30 % of the population has only a basic education level, 50 % has secondary education level and the remaining 20 % has superior education level. the description can be given in the form of a modal object as:

[Gender ~ {male (0.5), female (0.5)}] $\wedge$

[Education ~ {basic (0.3), secondary (0.5), superior (0.2)}]

In this case, "Gender" and "Education" are modal variables, '~' expresses that the frequency distributions on the categories are exactly those given, and '$\wedge$' represents a conjunction.

Each individual $\omega \in E$ is described by a modal object:

$$s(\omega) = \bigwedge_{i=1}^{p} [Y_i(\omega) \in \{y_1^i(p_1^i(\omega)),\cdots, y_{k_i}^i(p_{k_i}^i(\omega))\}]$$

In the description of each individual $\omega \in E$, we always have $p_1^i(\omega) + \cdots + p_{k_i}^i(\omega) = 1$, the associated objects are hence called "probabilistic". For sake of simplicity, we will also call "probabilistic" an object for which this condition is not imposed. In fact, when $p_1 + p_2 + \ldots + p_k = 1$, we are in presence of a probability or frequency distribution, while for $p_1 + p_2 + \ldots + p_k \geq 1$ *(resp.* $p_1 + p_2 + \ldots + p_k \leq 1$) we have an upper cover (resp. lower cover) of a probability or frequency distribution.

DEFINITION 3

We define a partial order relation in the set of probabilistic objects defined on the same variable set $\{Y_1, \ldots, Y_p\}$, as follows:

If $\quad s_1 = \bigwedge_{i=1}^{p} [Y_i\ R\ \{y_1^i(p_1^i),\cdots, y_{k_i}^i(p_{k_i}^i)\}] \qquad s_2 = \bigwedge_{i=1}^{p} [Y_i\ R\ \{y_1^i(q_1^i),\cdots, y_{k_i}^i(q_{k_i}^i)\}]$

then $s_1 \subseteq s_2$ iff $p_j^i \leq q_j^i$, j=1,…,k_i, i=1,..., p.

EXAMPLE 3. Consider again the previous example. Then:

[Gender R {male (0.4), female (0.6)}] $\subseteq$ [Gender ~R {male (0.5), female (0.7)}] .

3. GALOIS CONNECTIONS

Let us start by recalling some notions, which will be used in the sequel.

DEFINITION 4

Let A be a set and $P(A)$ the power set of A. A **closure operator** on A is a mapping $h: P(A) \rightarrow P(A)$ which is extensive, idempotent and isotone, that is:

$X \subseteq h(X)$ (extensivity)
$h(X) = h(h(X))$ (idempotence)
$X \subseteq Y \Rightarrow h(X) \subseteq h(Y)$ (isotony)

A subset X of A is said to be **closed** if $X = h(X)$.

DEFINITION 5

An **anti-closure operator** on a set A is a mapping $h : P(A) \rightarrow P(A)$ which is anti-extensive, idempotent and isotone, that is:

$h(X) \subseteq X$ (anti-extensivity)
$h(X) = h(h(X))$ (idempotence)
$X \subseteq Y \Rightarrow h(X) \subseteq h(Y)$ (isotony)

A subset X of A is said to be **open** if $X = h(X)$.

DEFINITION 6

Let $(A, \leq_1)$ and $(B, \leq_2)$ be two ordered sets.

A **Galois connection** is a pair (f,g), where f is a mapping $f: A \rightarrow B$, g is a mapping $g: B \rightarrow A$, such that f and g are antitone, and $h=gof$ and $h'=fog$ are extensive. Formally,

$x \leq_1 x_1 \Rightarrow f(x) \geq_2 f(x_1)$
$y \leq_2 y_1 \Rightarrow g(y) \geq_1 g(y_1)$
for any $x \in A, y \in B, x \leq_1 g(f(x))$ and $y \leq_2 f(g(y))$

h and h' are closure operators.

DEFINITION 7

A **lattice** is a partial ordered set A such that for any two elements there is a join (least upper bound) and a meet (greatest lower bound) in A. A lattice is said to be complete if any subset of A has a join and a meet in A.

As an example, the order relation introduced in Definition 3, allows defining a lattice on the set of probabilistic objects. In this lattice, the join and meet of a pair of objects

$$s_1 = \prod_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(r_1^i), \cdots, y_{k_i}^i(r_{k_i}^i) \right\} \right] \quad \text{and} \quad s_2 = \prod_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(q_1^i), \cdots, y_{k_i}^i(q_{k_i}^i) \right\} \right],$$

are given, respectively, by:

$$s_1 \sqcup s_2 = \prod_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i) \right\} \right], \text{ with } t_j^i = \text{Max}\{r_j^i, q_j^i\}, j=1,\ldots,k_i, i=1,\ldots,p$$

$$s_1 \sqcap s_2 = \prod_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(z_1^i), \cdots, y_{k_i}^i(z_{k_i}^i) \right\} \right], \text{ with } z_j^i = \text{Min}\{r_j^i, q_j^i\}, j=1,\ldots,k_i, i=1,\ldots,p$$

DEFINITION 8 [Leclerc, 1994]

Let $(P, \leq)$ be an ordered set. An element $x \in P$ is said to be **join-irreducible** if it is not the join of a finite part of P not containing it. Dually, we define a **meet-irreducible** element.

The notion of irreducible elements may be used to simplify the lattice graphical representation [Duquenne, 1987; Godin *et al*, 1995].

Now, let S be the set of all probabilistic objects, with $0 \leq p_j \leq 1$.

THEOREM 1

The couple of mappings

$$f : S \to P(E)$$
$$s \to ext_E\, s = \{ \omega \in E : s(\omega) \geq s \}$$
$$g : P(E) \to S$$
$$\{ \omega_1, \ldots, \omega_k \} \to s = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i) \right\} \right]$$

with $t_j^i = \text{Max}\, \{p_j^i(\omega_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$

form a Galois connection between $(P(E), \subseteq)$ and $(S, \geq)$.

Proof. Due to the duality on S, we have to prove that, as mappings between $(P(E), \subseteq)$ and $(S, \subseteq)$, f and g are isotone, $h=gof$ is anti-extensive and $h'=fog$ is extensive.

a) f is isotone

Let $s_1 = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(p_1^i), \cdots, y_{k_i}^i(p_{k_i}^i) \right\} \right]$, $s_2 = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(q_1^i), \cdots, y_{k_i}^i(q_{k_i}^i) \right\} \right]$

and suppose that $s_1 \geq s_2$, i.e., $p_j^i \geq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$. Let $\omega \in f(s_1)$, i.e., $s(\omega) \geq s_1$.
Then we also have $s(\omega) \geq s_2$, that is, $\omega \in f(s_2)$. So, $f(s_1) \subseteq f(s_2)$, that is, f is isotone.

b) g is isotone

Let $A \subseteq B \subseteq E$, with no loss of generality we may write $A = \{\omega_1,\ldots,\omega_k\}$, $B = \{\omega_1,\ldots,\omega_k, \omega_{k+1}, \ldots, \omega_m\}$.

$g(A) = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i) \right\} \right]$, with $t_j^i = \text{Max}\, \{p_j^i(\omega_h), h=1,...,k\}$, $j=1,...,k_i$, $i=1,...,p$. Let $z_j^i = \text{Max}\, \{p_j^i(\omega_h), h=1,\ldots,m\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$, $g(B) = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(z_1^i), \cdots, y_{k_i}^i(z_{k_i}^i) \right\} \right]$.

Then, $t_j^i \geq z_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, and so $g(A) \geq g(B)$. That is, g is isotone.

c) h is anti-extensive

Let $s = \bigwedge_{i=1}^{p} \left[Y_i \sim \left\{ y_1^i(q_1^i), \cdots, y_{k_i}^i(q_{k_i}^i) \right\} \right]$,

$f(s) = \{\omega_1, \ldots, \omega_k\} \in P(E)$ such that $s(\omega_h) \geq s$, $h = 1,\ldots, k$,

where $s(\omega_h) = \bigwedge_{i=1}^{p} \left[Y_i(\omega_h) \sim \left\{ y_1^i(p_1^i(\omega_h)), \cdots, y_{k_i}^i(p_{k_i}^i(\omega_h)) \right\} \right]$,

and $p_j^i(\omega_h) \geq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, $h=1,\ldots,k$.

Now, $h(s) = g(f(s)) = \bigcap_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i)\}]$, with $t_j^i = \text{Max}\{p_j^i(\omega_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$. It follows that $t_j^i \geq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$., i.e., $h(s) \geq s$.

d) h' is extensive

Let $A \subseteq P(E)$, $A = \{\omega_1, \ldots, \omega_k\}$. $g(A) = s = \bigcap_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i)\}]$, with

$t_j^i = \text{Max}\{p_j^i(\omega_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$.

$h'(A) = f(g(A)) = \{\omega \in E : s(\omega) \geq s\}$

where $s(\omega) = \bigcap_{i=1}^{p} [Y_i(\omega) \sim \{y_1^i(p_1^i(\omega)), \cdots, y_{k_i}^i(p_{k_i}^i(\omega))\}]$, that is, ω such that $p_j^i(\omega) \geq t_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, $h=1,\ldots,k$. Since $p_j^i(\omega_h) \leq \text{Max}\{p_j^i(\omega_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$, it follows that $\omega_h \in h'(A)$, $\forall \omega_h \in A$, that is, $A \subseteq h'(A)$, h' is extensive.

Then (f,g) is a Galois connection between $(P(E), \subseteq)$ and $(S, \geq)$. In other terms, g is a so-called residuated mapping and f its associated residual mapping. It follows that h' is a closure operator and h an anticlosure operator [Leclerc, 1990].

EXAMPLE 4. Consider the following data array where four groups of people, $\omega 1$, $\omega 2$, $\omega 3$ and $\omega 4$ are described according to the distributions of variables Gender and Instruction:

	Gender	Instruction
$\omega 1$	Male (0.4), Fem.(0.6)	Prim.(0.3), Sec.(0.4), Sup.(0.3)
$\omega 2$	Male (0.1), Fem.(0.9)	Prim.(0.1), Sec.(0.2), Sup.(0.7)
$\omega 3$	Male (0.8), Fem.(0.2)	Prim.(0.2), Sec.(0.3), Sup.(0.5)
$\omega 4$	Male (0.5), Fem.(0.5)	Prim.(0.3), Sec.(0.2), Sup.(0.5)

Let $A = \{\omega 1, \omega 2\}$.
$g(A) = [\text{Gender} \in \{\text{Male}(0.4), \text{Fem.}(0.9)\}] \cap [\text{Instruction} \in \{\text{Prim.}(0.3), \text{Sec.}(0.4), \text{Sup.}(0.7)\}]$
$f(g(A)) = \{\omega 1, \omega 2\}$.

THEOREM 2

The couple of mappings

$f : S \rightarrow P(E)$
$s \rightarrow ext_E\, s = \{\omega \in E : s(\omega) \geq s\}$
$g : P(E) \rightarrow S$

$\{\omega_1, \ldots, \omega_k\} \rightarrow s = \bigcap_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i)\}]$

with $t_j^i = \text{Min}\{p_j^i(\omega_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$

form a Galois connection between $(P(E), \subseteq)$ and $(S, \leq)$.

Proof. We must prove that f and g are antitone and that $h=gof$ and $h'=fog$ are extensive.

a) f is antitone

Let $s_1 = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(p_1^i),\cdots,y_{k_i}^i(p_{k_i}^i)\}]$, $s_2 = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(q_1^i),\cdots,y_{k_i}^i(q_{k_i}^i)\}]$

and suppose that $s_1 \sqsubseteq s_2$, i.e., $p_j^i \sqsubseteq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$. Let $\Delta \in f(s_2)$, i.e., $s(\Delta) \geq s_2$.

Then we also have $s(\Delta) \geq s_1$, that is, $\Delta \in f(s_1)$. So, $f(s_2) \subseteq f(s_1)$, that is, f is antitone.

b) g is antitone

Let $A \subseteq B \subseteq E$, with no loss of generality we may write $A = \{\Delta_1, \ldots, \Delta_k\}$, $B = \{\Delta_1,\ldots,\Delta_k, \Delta_{k+1}, \ldots, \Delta_m\}$.

$g(A) = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i),\cdots,y_{k_i}^i(t_{k_i}^i)\}]$, with $t_j^i = \text{Min}\ \{p_j^i(\Delta_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$. Let $z_j^i = \text{Min}\ \{p_j^i(\Delta_h), h=1,\ldots,m\}$, $j=1,\ldots,k_i$, $i=1,...,p$, $g(B) = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(z_1^i),\cdots,y_{k_i}^i(z_{k_i}^i)\}]$.

Then, $z_j^i \sqsubseteq t_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, and so $g(B) \subseteq g(A)$. That is, g is antitone.

c) h is extensive

Let $s = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(q_1^i),\cdots,y_{k_i}^i(q_{k_i}^i)\}]$,

$f(s) = \{\Delta_1, \ldots, \Delta_k\} \subseteq P(E)$ such that $s(\Delta_h) \geq s$, $h = 1,\ldots, k$,

where $s(\Delta_h) = \bigwedge_{i=1}^{p} [Y_i(\Delta_h) \sim \{y_1^i(p_1^i(\Delta_h)),\cdots,y_{k_i}^i(p_{k_i}^i(\Delta_h))\}]$,

and $p_j^i(\Delta_h) \geq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, $h=1,\ldots,k$.

Now, $h(s) = g(f(s)) = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i),\cdots,y_{k_i}^i(t_{k_i}^i)\}]$, with $t_j^i = \text{Min}\{p_j^i(\Delta_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$. It follows that $t_j^i \geq q_j^i$, $j=1,\ldots,k_i$, $i=1,...,p$, i.e., $s \sqsubseteq h(s)$.

d) h' is extensive

Let $A \subseteq P(E)$, $A = \{\Delta_1, \ldots, \Delta_k\}$. $g(A) = s = \bigwedge_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i),\cdots,y_{k_i}^i(t_{k_i}^i)\}]$, with

$t_j^i = \text{Min}\ \{p_j^i(\Delta_h), h=1,...,k\}$, $j=1,\ldots,k_i$, $i=1,...,p$.

$h'(A) = f(g(A)) = \{\Delta \in E : s(\Delta) \geq s\}$

where $s(\omega) = \bigcap_{i=1}^{p} [Y_i(\omega) \sim \{y_1^i(p_1^i(\omega)), \cdots, y_{k_i}^i(p_{k_i}^i(\omega))\}]$, that is ω such that $p_j^i(\omega) \geq t_j^i$, $j=1,\ldots,k_i$, $i=1,\ldots,p$, $h=1,\ldots,k$. Since $p_j^i(\omega_h) \geq \text{Min}\ \{p_j^i(\omega_h),\ h=1,\ldots,k\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$, it follows that $\omega_h \in h'(A)$, $\forall \omega_h \in A$, that is, $A \subseteq h'(A)$, h' is extensive.

Then (f,g) is a Galois connection between $(P(E), \subseteq)$ and $(S, \leq)$. It follows that h and h' are closure operators.

EXAMPLE 5. Consider again the data array of Example 4, and let $B = \{\omega 2, \omega 3\}$.
$g(B) = [\text{Gender} \in \{\text{Male}(0.1), \text{Fem.}(0.2)\}] \wedge [\text{Instruction} \in \{\text{Prim.}(0.1), \text{Sec.}(0.2), \text{Sup.}(0.5)\}]$
$f(g(B)) = \{\omega 2, \omega 3, \omega 4\}$.

In fact, Theorems 1 and 2 constitute a special case of a general framework, which could be summarised as follows: Let $P(E)$ be the finite power set of a set E and Q a set of descriptions, ordered by generalisation and such that each element $x \in E$ has a description $d(x) \in Q$. If two mappings f and g are defined, where f associates to each element $q \in Q$ the set of elements of E that verify the description q, and g associates to each element $A \in P(E)$ the least general description verified by all elements of A, then (f,g) constitutes a Galois connection between Q and $P(E)$.

Theorem 1 is using as description space the upper cover of frequency or probability distributions, while Theorem 2 is using as description space the lower cover of frequency or probability distributions. In fact, we are using Choquet capacity or credibility type distributions as description spaces [Diday, Emilion, 1997].

DEFINITION 9

A probabilistic object $s \in S$ is said to be **complete** is $h(s) = s$.

PROPOSITION 1 [Brito, 1991]

Let s be a complete probabilistic object and A its extent, $A = f(s)$. Then, $g(A) = s$ and $A = h'(A)$. Conversely, if $A = h'(A)$ and $s = g(A)$, then s is complete and $A = f(s)$.

DEFINITION 10

Given a set of observed objects, E, a **concept** is defined as a pair (A, s), where $A \subseteq E$, $s \in S$, s is complete and $A = f(s)$.

EXAMPLE 6. Consider again Examples 4 and 5. Then $(A, g(A))$ is a concept whereas $(B, g(B))$ is not.

Since we have proved in Theorems 1 and 2 that we obtain Galois connections between two lattices, we get as an immediate consequence Theorems 3 and 4 below, that we state without proof [Birkhoff, 1967; Barbut, Monjardet, 1970].

THEOREM 3

Let (f_1,g_1) be the Galois connection in theorem 1.

If $s_1 = \bigcap_{i=1}^{p} [Y_i \sim \{y_1^i(r_1^i), \cdots, y_{k_i}^i(r_{k_i}^i)\}]$ and $s_2 = \bigcap_{i=1}^{p} [Y_i \sim \{y_1^i(q_1^i), \cdots, y_{k_i}^i(q_{k_i}^i)\}]$

we define $s_1 \sqcap s_2 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i)\}]$, with $t_j^i = \text{Max}\ \{r_j^i, q_j^i\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$ and $s_1 \sqcup s_2 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(z_1^i), \cdots, y_{k_i}^i(z_{k_i}^i)\}]$, with $z_j^i = \text{Min}\ \{r_j^i, q_j^i\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$.

Then the set of concepts, ordered by $(A_1, s_1) \sqsubseteq (A_2, s_2) \Leftrightarrow A_1 \subseteq A_2$ is a lattice where meet and join are given by:

$$\inf((A_1, s_1), (A_2, s_2)) = (A_1 \cap A_2, (g_1 \circ f_1)(s_1 \sqcup s_2))$$
$$\sup((A_1, s_1), (A_2, s_2)) = ((f_1 \circ g_1)(A_1 \cup A_2), s_1 \sqcap s_2)$$

In the sequel, this lattice will be called "union lattice".

THEOREM 4

Let (f_2, g_2) be the Galois connection in theorem 2.

If $s_1 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(r_1^i), \cdots, y_{k_i}^i(r_{k_i}^i)\}]$ and $s_2 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(q_1^i), \cdots, y_{k_i}^i(q_{k_i}^i)\}]$

we define $s_1 \sqcap s_2 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(t_1^i), \cdots, y_{k_i}^i(t_{k_i}^i)\}]$, with $t_j^i = \text{Min}\ \{r_j^i, q_j^i\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$ and $s_1 \sqcup s_2 = \prod_{i=1}^{p} [Y_i \sim \{y_1^i(z_1^i), \cdots, y_{k_i}^i(z_{k_i}^i)\}]$, with $z_j^i = \text{Max}\ \{r_j^i, q_j^i\}$, $j=1,\ldots,k_i$, $i=1,\ldots,p$.

Then the set of concepts, ordered by $(A_1, s_1) \sqsubseteq (A_2, s_2) \Leftrightarrow A_1 \subseteq A_2$ is a lattice where meet and join are given by:

$$\inf((A_1, s_1), (A_2, s_2)) = (A_1 \cap A_2, (g_2 \circ f_2)(s_1 \sqcup s_2))$$
$$\sup((A_1, s_1), (A_2, s_2)) = ((f_2 \circ g_2)(A_1 \cup A_2), s_1 \sqcap s_2)$$

In the sequel, this lattice will be called "intersection lattice".

In this section we have proved that Galois lattices can be defined on probabilistic data. In the next section, we give an algorithm which allows obtaining the elements of these Galois lattices. In Section 5 we apply the algorithm to obtain Galois lattices on real data.

4. ALGORITHMS

When considering binary data or discrete data, Ganter's algorithm [Ganter, Wille, 1999] searches for closed sets by exploring the closure space in a certain order. When considering large databases, we have chosen this algorithm, because it requires little memory size at each step, by considering only one concept when looking for the next one. We have implemented this algorithm using the Galois connections defined above on probabilistic data. To reduce computational complexity, we look for the concepts exploring the set of individuals. In fact, when treating concepts, it is equivalent to search for closed sets on individuals or on variables.

The description of Ganter's algorithm can be found for instance in [Kuznetsov, Obiedkov, 2002], where its performance is compared with other algorithms.

Theorems 3 and 4 tell us that the set of concepts constitutes a Galois lattice. Ganter's algorithm provides all concepts which can be found in the data set, by enumerating vectors of individuals. Implicitly, with vectors of individuals and Galois connections, we can obtain all concepts (A,s) of the Galois Lattice [Guénoche, 1990]. So, by applying this algorithm directly on probabilistic data and using the Galois connections defined in the previous section, we obtain all possible concepts and they form a Galois lattice.

We now present an algorithm to construct the lattice. In fact, Ganter's algorithm enumerates the concepts, but it does not construct the lattice.

LATTICE CONSTRUCTION

The graphical representation of the Galois lattice, named Hasse diagram, is a graph where the vertex set is the set of concepts and the edges represent the covering relation of the quasi-order relation.

In a recent paper, Kuznetsov and Obiedkov [2002] compare the performance of algorithms for generating concept lattices. They present several classical and new algorithms and extend each classical algorithm adding a drawing diagram step. Considering that:

- the performance of the algorithms depends on the type of the data,
- probabilistic data are more complex than the data considered in [Kuznetsov, Obiedkov, 2002],
- many of the algorithms are very complex to extend to probabilistic data,
- for large and dense datasets, the fastest algorithms are those proposed by [Ganter, 1999; Norris, 1978],

we choose to use Ganter's algorithm conjointly with the following drawing diagram algorithm.

We use the graph terminology to define our algorithm. A graph is a couple formed by a set of vertices and a set of edges. In an oriented graph, the edges have a unique sense. We define a path between two vertices by a list of vertices where two successive vertices are joined with a directed edge. We denote by root of an oriented graph a vertex where no path ends.

Our aim is to design an oriented graph with the supremum of the Galois Lattice as root. A directed edge will refer to the inclusion order: if we have two objects such that $s_1 \subset s_2$, s_2 will be the origin (also designed by predecessor) of the directed edge of these two vertices and s_1 the destination (also designed by successor).

A naïve approach is for each object s to look for the included objects s_i by testing for each one if there is no object s_j such that $s_i \subset s_j \subset s$. The complexity is $o(N^3)$ where N=card(E).

In order to optimise this search for large data sets, we propose an algorithm which is an extension of the one proposed by [Sedgewick, 1999] in the context of graph

exploration. Here, this extension allows us to add a node and explore the graph to discover links simultaneously with a minimum number of explorations.

The general principle is the following: we have two lists: $C=(c_1,\ldots,c_k)$ is the concept list obtained by Ganter's algorithm and $R=(r_1,\ldots, r_l)$ the root list obtained during this algorithm. At the end of the algorithm, this root list contains only one element: the supremum of the lattice. For each concept c_i, and for each root r_j, we look if r_j is included in c_i by exploring the existing graph from r_j and c_i is added to the root list.

Formally, we have:

```
The concept list contains all concepts of the Galois lattice
The root list is empty
For each concept C in the concept list
  For each root R in the root list
    Set a colour depending on C
    explore(C, R, colour)
    add C in the root list
```

The principle of the exploration part is the following: if r_j is included in c_i, we create the node c_i and an edge between c_i and r_j and mark all the vertices on a path from r_j as linked, else we explore the successors of r_j and recursively look for inclusion. During the process, we stop the exploration if we find a linked vertex. We use an attribute colour to assure that for c_i, we are going to explore each node of the existing graph only once. Then, we look in this graph for included concepts only if they are not yet coloured by c_i.

Formally, we have:

```
if R is not coloured by colour
  if R ⊆ C
    add the node C to the graph and an edge between C and R
    Colour R with colour
    Suppress R from root list
    for each recursively successors S of R
      Colour S with colour
      if an edge between C and S exists then suppress it
  else if R ⊃ C
    for each successor S of R
      explore(C, S, colour)
```

When we insert a new concept in the graph, we explore all nodes of the existing graph only once. So the complexity of this algorithm is o(*N*_).

Finally, we obtain a graph, where the vertices are the concepts, which are described both by intent and extent, and the edges represent the partial order relation between concepts (see Definition 3). With these specifications, it is quite easy to represent graphically the Galois lattice.

The search of all concepts of the Galois lattice and its construction with this algorithm will be applied in the next section.

5. APPLICATIONS

In this section, we first compare the lattices obtained by the algorithms described above to those obtained by other authors, on the basis of published examples. On the compared examples with small data sets, we present the concepts extracted from both the intersection Galois lattice and the union Galois lattice. Hence we can see that the two lattices are

complementary and present quite the same information, considering two different points of view. We further apply these algorithms to two real data sets; the first of these applications illustrates the construction of the intersection Galois lattice, and the second application illustrates the construction of the union Galois lattice. In all examples, the mapping *f*, which defines the extent of a probabilistic object, is not explicitly indicated, it is given by the relation *R* used in the complete objects that constitute the intent of the obtained concepts.

5.1 COMPARISON WITH OTHER WORKS

The algorithms described in Section 4 have been applied to data sets used by Burusco and Fuentes-Gonzales [1998] and Herrmann, Hölldobler and Strohmaier [1996].

5.1.1 *Burusco and Fuentes-Gonzales data*

The data set presented in (Burusco and Fuentes-Gonzales, 1998) describes three illnesses Illness1, Illness2 and Illness3 on the basis of the presence of three symptoms Symptom1, Symptom2 and Symptom3; the data matrix registers the L-fuzzy relation as follows:

	Illness1	Illness2	Illness3
Symptom1	0.6	0.6	0
Symptom2	0.9	0.5	0.3
Symptom3	1	0.2	0.7

Figure 1 represents the union-lattice obtained on this data set.

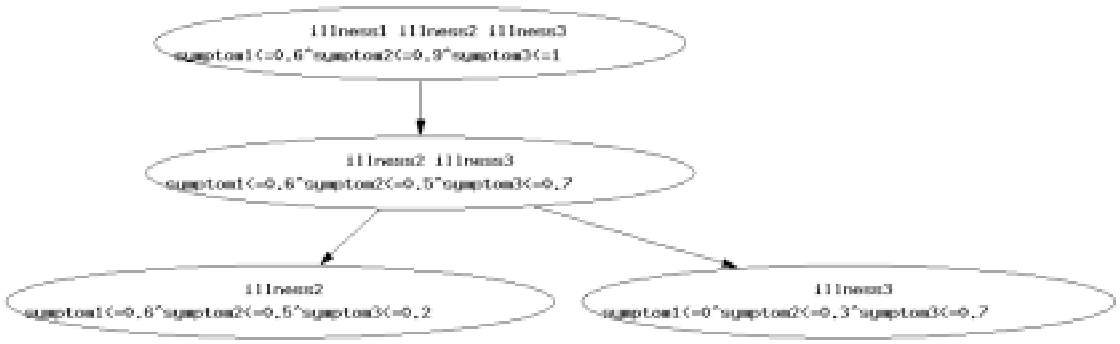

Figure 1. Union concept lattice of Burusco and Fuentes-Gonzales data.

The corresponding concepts are:
({Illness2}, $a1$) with
$a1$ = [Symptoms ⊇ {Symptom1(0.60), Symptom2(0.50), Symptom3(0.20)}]
({Illness3}, $a2$) with
$a2$ = [Symptom1 ⊇ {Symptom1(0.00), Symptom2(0.30), Symptom3(0.70)}]
({Illness2, Illness3}, $a3$) with
$a3$ = [Symptoms ⊇ {Symptom1(0.60), Symptom2(0.50), Symptom3(0.70)}]
({Illness1, Illness2, Illness3}, $a4$) with
$a4$ = [Symptoms ⊇ {Symptom1(0.60), Symptom2(0.90), Symptom3(1.00)}]

Figure 2 represents the intersection-lattice obtained on this data set. The corresponding concepts are:
({Illness1}, $b1$) with
$b1$ = [Symptoms ≥ {Symptom1(0.60), Symptom2(0.90), Symptom3(1.00)}]
({Illness1, Illness2}, $b2$) with
$b2$ = [Symptoms ≥ {Symptom1(0.60), Symptom2(0.50), Symptom3(0.20)}]
({Illness1, Illness3}, $b3$) with
$b3$ = [Symptoms ≥ {Symptom1(0.00), Symptom2(0.30), Symptom3(0.70)}]
({Illness1, Illness2, Illness3}, $b4$) with
$b4$ = [Symptoms ≥ {Symptom1(0.00), Symptom2(0.30), Symptom3(0.20)}]

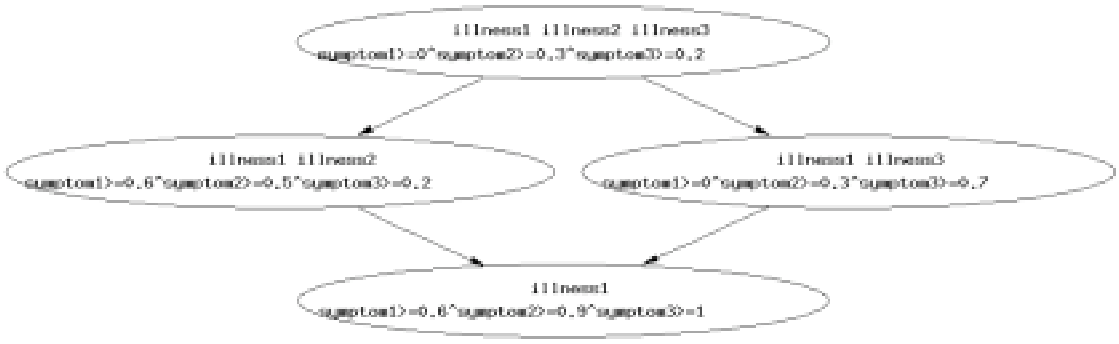

Figure 2. Intersection concept lattice of Burusco and Fuentes-Gonzales data.

In both cases, our algorithm produces much less concepts than those proposed by Burusco's method. It should be noticed that in our case, the membership of individuals (here, the illnesses) to the concepts is deterministic, whereas in the authors' concepts it is fuzzy. Moreover, in our case, the values associated to categories are either minimum (for the intersection lattice) or maximum (for the union lattice) values observed in the corresponding extent set, and not fuzzy relations. As a consequence, the concepts cannot be directly compared, nevertheless there seems to be no incongruence between both results. For instance, the interpretation of concept 8 in [Burusco, Fuentes-Gonzales, 1998, p. 113] can be found by looking at concepts ({Illness1, Illness2}, b2) and ({Illness1, Illness3}, $b3$) in our intersection lattice.

5.1.2 *Herrmann, Hölldobler and Strohmaier data*

The data set presented in (Herrmann, Hölldobler and Strohmaier, 1996) describes a group of persons in a hypertension patient record database, on the basis of three fuzzy variables, Tinnitus, Headache and Blood Pressure:

	TINNITUS		HEADACHE		BLOOD PRESSURE		
Name	Often	Seldom	Often	Seldom	High	Normal	Low
Ann	0.8	0.2	0.9	0.1	0.8	0.2	0.0
Bob	1.0	0.0	0.0	1.0	0.6	0.4	0.0
Chris	1.0	0.0	0.1	0.9	0.9	0.1	0.0
Doug	0.3	0.7	0.7	0.3	0.0	0.6	0.4
Eve	0.6	0.4	0.7	0.3	0.0	0.8	0.2

By applying our algorithm to this data set, we obtain, in both the union and the intersection lattice, more concepts that the authors: 32 concepts in the union concept lattice and 25 concepts in the intersection concept lattice, while the authors obtain only 7. Taking into account the differences in the concepts' description language, the authors' concepts may be found among ours. For instance, concept:

c : {{Ann, Bob, Chris}{Tinnitus (often), Blood Pressure (high)}, □=0.85, □=0.11}

roughly corresponds to the following concept of the union concept lattice:

({Ann, Bob, Chris}, $a7^u$) with

$a7^u$ = [Tinnitus □ {often(1.0), seldom(0.2)}] ^
[Headache □ {often(0.9), seldom(1.0)}] ^
[Blood Pressure □ {high(0.9), normal (0.4), low(0.0)}]

and to the following concept of the intersection concept lattice:

({Ann, Bob, Chris}, $a7^i$) with

$a7^i$ = [Tinnitus ≥ {often(0.8), seldom(0.0)}] ^
[Headache ≥{often(0.0), seldom(0.1)}] ^
[Blood Pressure ≥ {high(0.6), normal (0.1), low(0.0)}]

It should be noticed, however, that whereas Herrmann, Hölldobler and Strohmaier consider mean values for each category in the concepts description, we consider either minimum (for the intersection lattice) or maximum (for the union lattice) values observed in the corresponding extent set.

5.2 CULTURAL SURVEY DATA

This data set has been obtained from a survey, made in 5 portuguese towns in 1997, on attendance of cultural and leisure activities. The 1409 individuals in the survey database have been aggregated according to professional activity, leading to a data set of 11 probabilistic objects describing the distribution of each variable in each professional group. Among the large number of variables in the base, we have chosen: instruction, with three categories - primary, secondary or superior - and football matches' attendance with three categories - yes, no, no answer - leading to the following data table:

	GROUP	INSTRUCTION	FOOTBALL
1	Students	Pr(0.08), Sec(0.88), Sup(0.03)	yes (0.51), no(0.49), no_ans (0.01)
2	Retired	Pr(0.92), Sec(0.05), Sup(0.03)	yes (0.12), no(0.88)
3	Employed	Pr(0.59), Sec(0.39), Sup(0.02)	yes (0.22), no(0.78)
4	Small independents	Pr(0.62), Sec(0.31), Sup(0.07)	yes (0.32), no(0.67), no_ans (0.01)
5	Housewives	Pr(0.93), Sec(0.07)	yes (0.10), no(0.90)
6	Intermediate executives	Pr(0.17), Sec(0.50), Sup(0.33)	yes (0.26), no(0.73), no_ans (0.01)
7	Industrial workers	Pr(0.73), Sec(0.27), Sup(0.01)	yes (0.40), no(0.60)
8	Intellectual/Scientific Executives	Sec(0.01), Sup(0.99)	yes (0.22), no(0.78)
9	Other	Pr(0.72), Sec(0.229), Sup(0.07)	yes (0.19), no(0.80), no_ans (0.01)
10	Manager / Independent Profession	Sec(0.02), Sup(0.98)	yes (0.28), no(0.70), no_ans (0.02)
11	Businessmen	Pr(0.08), Sec(0.33), Sup(0.58)	yes (0.42), no(0.58)

Applying the algorithms described in Section 4 to this data set, using the Galois connection obtained in Theorem 2 and considering the result of Theorem 4, we have obtained an intersection concept lattice on the 11 professional groups. Figure 3 represents this lattice, where each node is represented by the extent of the corresponding concept. Figure 4 represents the irreducible elements of the lattice (see Definition 8).

As discussed in the Introduction, a Galois lattice may be considered as a clustering, allowing us to identify concepts which are “interesting” for the user.

From the irreducible lattice on the Cultural Survey data, we have focused on two groups of concepts:

- the general concepts, located at the top of the lattice;
- the specific concepts, located at the bottom of the lattice.

We find that some categories are described by exactly the same values for all concepts. These categories are: 'primary' for the variable instruction and 'no answer' for the variable football. For all concepts, we have: instruction_primary$\geq$0.08 and football_no_answer$\geq$0. This means that all groups contain the same minimum of individuals with a primary instruction level and that the number of no answer for going to football matches has always a minimum of 0. These categories will not be taken into account in the following comparisons.

Figure 3. Intersection concept lattice of the culture data.

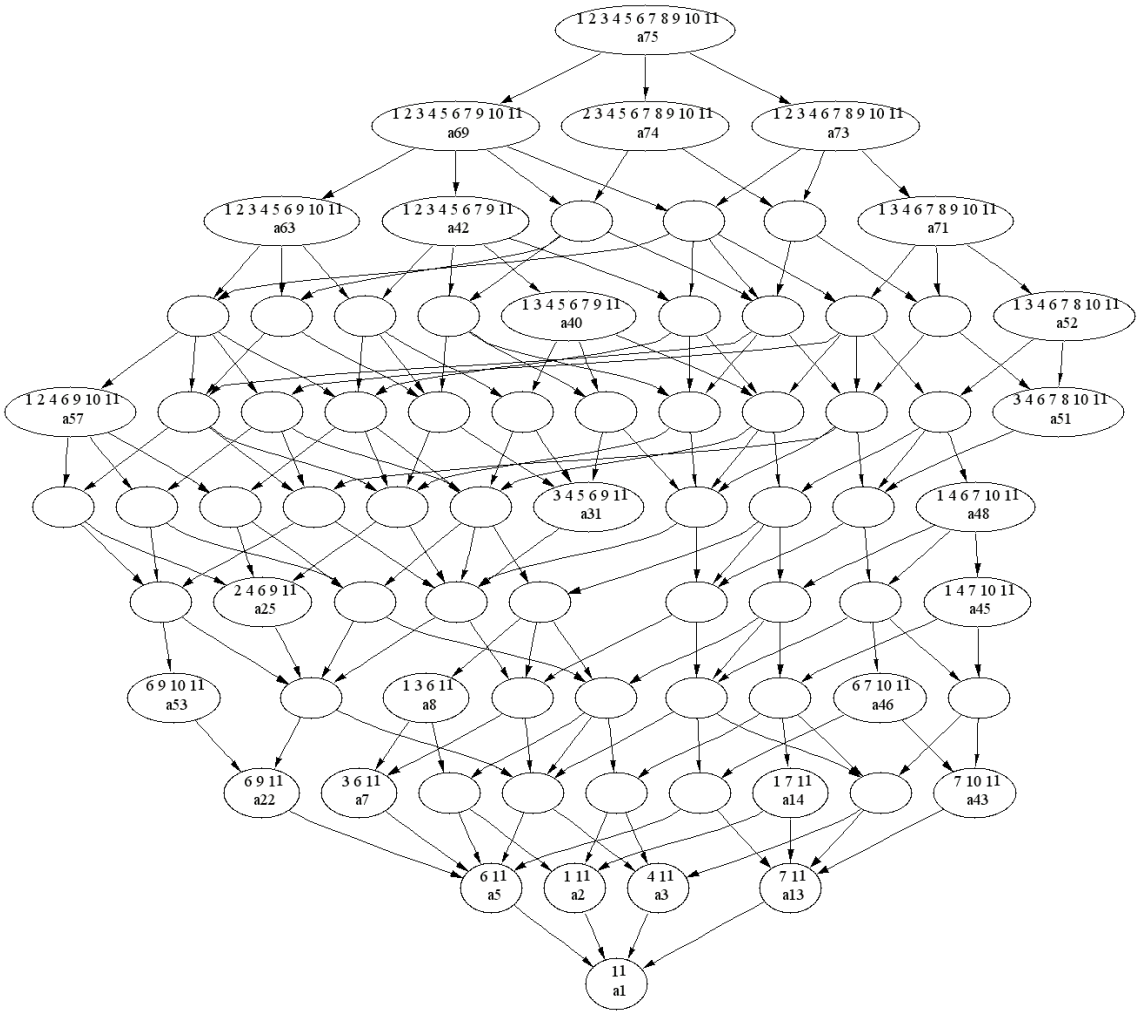

Figure 4. Irreducibles of the intersection concept lattice of the culture data.

The general concepts considered are:

c69 = ({1, 2, 3, 4, 5, 6, 7, 9, 10, 11}, a69) with
a69 = [Instruction ≥ {secondary(0.02), superior(0.01)}] ^
[Football ≥ {no(0.49) yes(0.10)}]

c73 = ({1, 2, 3, 4, 6, 7, 8, 9, 10, 11}, a73) with
a73 = [Instruction ≥ {secondary(0.01), superior(0.01)}] ^
[Football ≥ {no(0.49), yes(0.12)}]

c74 = ({2, 3, 4, 5, 6, 7, 8, 9, 10, 11}, a74) with
a74 =[Instruction ≥ {secondary(0.01), superior(0.01)}] ^
[Football ≥ {no(0.58), yes(0.10)}]

By comparison, we can notice that:

- All concepts contain the same minima of individuals with a superior level of instruction.
- Concept c69 doesn't contain the individuals '8- Intellectual/Scientific Executives' and has a minimum of a slightly higher level of secondary instruction.

- Concept c73 doesn't contain the individuals '5- Housewives' and has a minimum slightly higher of individuals going to football matches.
- Concept c74 doesn't contain the individuals '1-Students', and has a minimum slightly higher of individuals not going to football matches.

Three trends are emerging: one group with a higher value for secondary level of instruction, one group with more people going to football matches and another with more people not going to matches.

The specific concepts considered are:

c7 = ({3, 6, 11}, a7) with
a7 = [Instruction ≥ {secondary(0.33), superior(0.02)}] ^
[Football ≥ {no(0.58), yes(0.22)}

c14 =({1, 7, 11}, a14) with
a14 =[Instruction ≥ {secondary(0.26), superior(0.01)} ^
[Football ≥ {no(0.49), yes(0.40)}]

c22 =({6, 9, 11}, a22) with
a22 =[Instruction ≥ {secondary(0.22), superior(0.07)}] ^
[Football ≥ {no(0.58), yes(0.19)}]

c43 =({7, 10, 11}, a43) with
a43 =[Instruction ≥ {secondary(0.02), superior(0.01)] ^
[Football ≥ {no(0.58), yes(0.28)}]

By comparison, we can notice that:

- The individual '11-Businessmen' is in the extent of all concepts. He is also in the minimum concept of the lattice. This group of individuals is particular as all minimum values are greater than the ones in other groups. This could be interpreted as the responses given by this group are quite homogeneous.
- The concept c7 gathers '3-Employed' and '6-Intermediate executive'. It distinguishes itself by a high minimum of secondary instruction level.
- The concept c14 gathers '1-Student' and '7-Ind.Worker'. It distinguishes itself by a relatively high minimum of secondary instruction level and a high minimum of football_yes.
- The concept c22 gathers '6-Intermediate executive' and '9-Other'. It distinguishes itself by a high minimum of football_yes and a slightly high level of instruction_superior.
- The concept c43 gathers '7-Ind.Worker' and '10-Manager / Indep. Professional'. It distinguishes itself by a low minimum of instruction_secondary.
- The trends are that, on the one hand we have a high secondary instruction level and, on the other hand, a high number of individuals going to football matches. In between we find some groups between these two trends.

Conclusion: General trends and specific trends agree. On the one hand, we have groups with a secondary level of instruction, while on the other hand, we have groups who go to football matches. The irreducible lattice clearly represents this separation. However, we can notice that the order on the categories of the variable ‘Instruction’ has been lost.

5.3 EMPLOYMENT DATA

This data set has been obtained from a survey, made in Portugal in 1998 by the National Statistical Institute, on the employment situation. The 22660 individuals in the survey database have been aggregated according to sex and age group, leading to a data set of 12 probabilistic objects describing the distribution of each variable in the corresponding group. The following variables have been chosen:

- Marital status: Single (S), Married (M), Widow (W), Divorced(D);
- Education: Without education (No), Primary (Pr), Secondary (Sec), Superior (Un);
- Economic activity (CEA): agriculture, cattle, hunt, forestry & fishing (Ag), construction (Cr), other services (Ot), real estate, renting & business activities (Re), wholesale and retail trade, repairs (Wh), public administration (Pub), manufacturing (Man), transport, storage & communication (Tr), hotels & restaurants (Hot), electricity, gas & water (El), financial intermediation (Fi), mining & quarrying (Min);
- Profession: skilled agriculture and fishery workers (Sk), elementary occupations (El), plant and machine operators and assemblers (Pla), craft and related trade workers (Cr), professionals (Pr), clerks (Cl), service, shop & market sales workers (Serv), technicians and associate professionals (Tech), legislators, senior officers and managers (Lge), Armed forces (Arm);
- Searching employment: yes/no;
- Full/part time.
- The probabilistic data table is presented in the annex.

Applying the algorithms described in Section 4 to this data set, using the Galois connection obtained in Theorem 1 and considering the result of Theorem 3, we have obtained a union concept lattice on the 12 groups. Figure 5 represents this lattice, where each node is represented by the extent of the corresponding concept. Figure 6 represents the irreducible elements of this lattice. In both figures, the infimum of the lattice, whose extent is empty, is not represented.

We have chosen to detail the interpretation of concepts c1, c3, and c19 as specific concepts, and c18 and c30 as more general concepts.

c1 = ({Men 15-24, Men 25-34, Men 35-44, Men 55-64, Women 15-24, Women 25-34, Women 35-44, Women 45-54}, a1) with

a1 = [Status $\sqsubseteq$ {widow(0.06), divorced(0.14), married(0.95), single(0.71)}] ^
[CEA $\sqsubseteq$ {agriculture, cattle, hunt, forestry & fishing(0.26), construction(0.26), other services(0.34), real estate, renting & business activities(0.06), wholesale and retail trade, repairs(0.18), public administration(0.09), manufacturing(0.38), transport, storage & communication(0.06), hotels & restaurants(0.10), electricity, gas & water(0.02), financial intermediation(0.03), mining & quarrying(0.01)}] ^

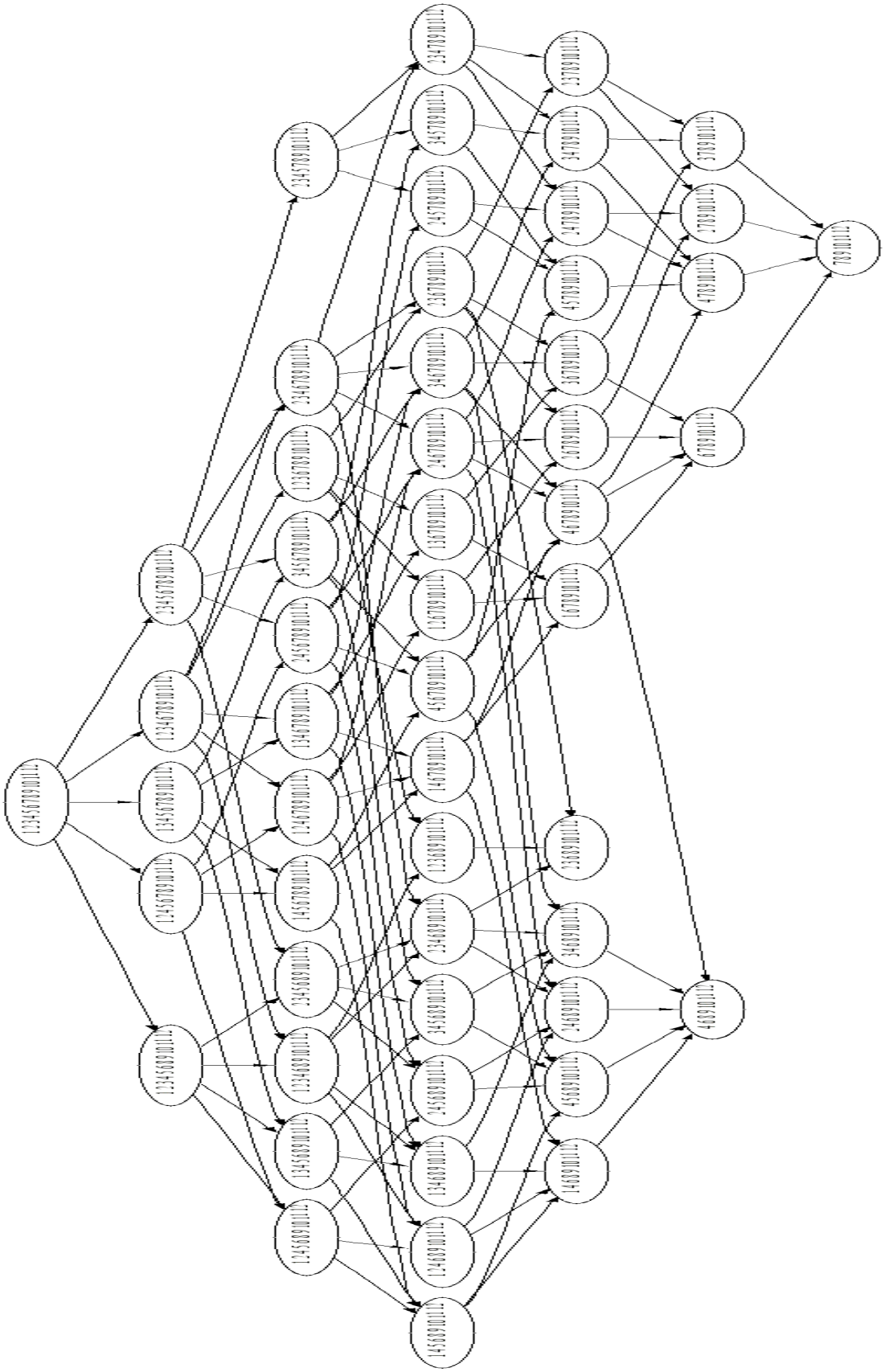

Figure 5. Union concept lattice of the employment data

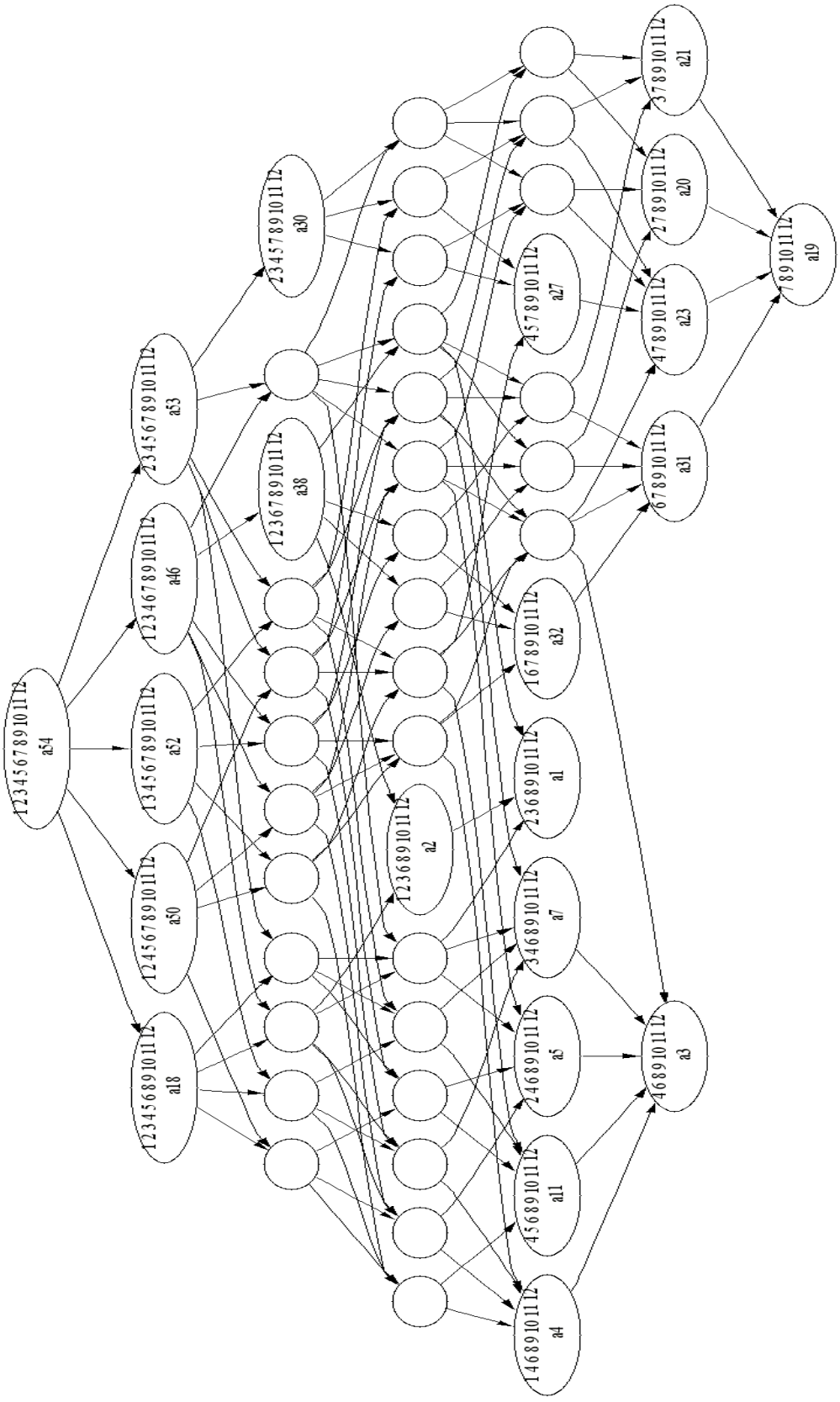

Figure 6. Irreducibles of the union concept lattice of the employment data

[Profession ∈ {skilled agriculture and fishery workers(0.23), elementary occupations(0.22), plant and machine operators and assemblers(0.13), craft and related trade workers(0.42), professionals(0.11), clerks(0.14), service, shop & market sales workers(0.28), technicians and associate professionals(0.10), legislators, senior officers and managers(0.14), armed forces(0.02)}] ^
[Education ∈ {without education(0.21), primary(0.65), secondary(0.42), superior(0.16)}] ^
[Searching ∈ {search_no(1.00), search_yes(0.04)}] ^
[Part/Full ∈ {part time(0.18), full time(0.98)}]

This class does not comprehend elderly people (more than 65 years old). This accounts for the lower value assigned to the category of "widow". Also, this class has a low value of category "part-time employed", "agriculture, cattle, hunt, forestry & fishing" as economic activity and "skilled agriculture and fishery workers" as profession.

c3 = ({Men 15-24, Men 25-34, Men 35-44, Men + 65, Women 15-24, Women 25-34, Women 35-44, Women 45-54}, a3) with

a3 = [Status ∈ {widow(0.10), divorced(0.14), married(0.90), single(0.71)}] ^
[CEA ∈ {agriculture, cattle, hunt, forestry & fishing (0.69), construction(0.26), other services(0.34), real estate, renting & business activities(0.06), wholesale and retail trade, repairs(0.18), public administration(0.09), manufacturing(0.38), transport, storage & communication(0.06), hotels & restaurants(0.10), electricity, gas & water(0.02), financial intermediation(0.02), mining & quarrying(0.01)}] ^
[Profession ∈ {skilled agriculture and fishery workers(0.65), elementary occupations(0.22), plant and machine operators and assemblers(0.12), craft and related trade workers(0.42), professionals(0.11), clerks(0.14), service, shop & market sales workers(0.28), technicians and associate professionals(0.10), legislators, senior officers and managers(0.10), armed forces(0.02)}] ^
[Education ∈ {without education(0.42), primary(0.64), secondary(0.42), superior(0.16)}] ^
[Searching ∈ {search_no(1.00), search_yes(0.04)}] ^
[Part/Full ∈ {part time(0.43), full time(0.98)}

In this class we find a high value for category "without education", and a quite high value of category "married". Otherwise the economic activity in "construction" has a rather low value, as has profession "legislators, senior officers and managers".

c19 = ({Men 15-24, Men 35-44, Men 45-54, Women 15-24, Women 25-34, Women 35-44}, a19) with

a19 = [Status ∈ {widow∈0.01 divorced∈0.14 married∈0.94 single∈0.71}] ^
[CEA ∈ {agriculture, cattle, hunt, forestry & fishing (0.10), construction(0.26), other services(0.34), real estate, renting & business activities(0.06), wholesale and retail trade, repairs(0.18), public administration(0.09), manufacturing(0.38), transport, storage & communication(0.08), hotels & restaurants(0.10), electricity, gas & water(0.02), financial intermediation(0.02), mining & quarrying(0.01)}] ^

[Profession □ {skilled agriculture and fishery workers(0.08), elementary occupations(0.19), plant and machine operators and assemblers(0.14), craft and related trade workers(0.42), professionals(0.11), clerks(0.14), service, shop & market sales workers(0.28), technicians and associate professionals(0.10), legislators, senior officers and managers(0.13),armed forces(0.02)}] ^
[Education □ {without education(0.05), primary(0.69), secondary(0.42), superior(0.16)}] ^
[Searching □ {search_no(0.99), search_yes(0.04)}] ^
[Part/Full □ {part time(0.11), full time(0.98)}]

In this class we do not find groups over 54 years old (for men) and 44 years old (for women). For this reason, the categories "without education" and "widow" present a very low value, and "primary education" a higher value. Also, we observe a low value for "part time worker" and "searching for a job", as well as for profession "skilled agriculture and fishery workers" and economic activity in "agriculture, cattle, hunt, forestry & fishing ".

c18 = ({Men 15-24, Men 25-34, Men 35-44, Men 55-64, Men + 65, Women 15-24, Women 25-34, Women 45-54, Women 35-44, Women 55-64, Women +65}, a18) with

a18 =[Status □ {widow(0.33), divorced(0.14), married(0.95), single(0.71)}] ^
[CEA □ {agriculture, cattle, hunt, forestry & fishing (0.70), construction(0.26), other services(0.34), real estate, renting & business activities(0.06), wholesale and retail trade, repairs(0.18), public administration(0.09), manufacturing(0.38), transport, storage & communication(0.06), hotels & restaurants(0.10), electricity, gas & water(0.02), financial intermediation(0.03), mining & quarrying(0.01)}] ^
[Profession □ {skilled agriculture and fishery workers(0.67), elementary occupations(0.28), plant and machine operators and assemblers(0.13), craft and related trade workers(0.42), professionals(0.11), clerks(0.14), service, shop & market sales workers(0.28), technicians and associate professionals(0.10), legislators, senior officers and managers(0.14), armed forces(0.02)}] ^
[Education □ {without education(0.70), primary(0.65), secondary(0.42), superior(0.16)}] ^
[Searching □ {search_no(1.00), search_yes(0.04)}] ^
[Part/Full □ {part time(0.61), full time(0.98)}]

In this class we observe a quite high value for status "widow" and for category "without education", perhaps associated to old people. Also, a high value for "part time worker" and for "agriculture, cattle, hunt, forestry & fishing" as economic activity.

c30 = ({Men 15-24, Men 25-34, Men 35-44, Men 45-54, Men 55-64, Men + 65, Women 15-24, Women 25-34, Women 35-44, Women +65}, a30) with

a30 =[Status □ {widow(0.33), divorced(0.14), married(0.95), single(0.71)}] ^
[CEA □ {agriculture, cattle, hunt, forestry & fishing (0.70), construction(0.26), other services(0.34), real estate, renting & business activities(0.06), wholesale and retail trade, repairs(0.18), public administration(0.09), manufacturing(0.38), transport, storage & communication(0.08), hotels &

restaurants(0.10), electricity, gas & water(0.02), financial intermediation (0.03), mining & quarrying (0.01)}] ^
[Profession □ {skilled agriculture and fishery workers(0.67), elementary occupations(0.19), plant and machine operators and assemblers(0.14), craft and related trade workers(0.42), professionals(0.11), clerks(0.14), service, shop & market sales workers(0.28), technicians and associate professionals(0.10), legislators, senior officers and managers(0.14), armed forces(0.02)}] ^
[Education □ {without education(0.70), primary(0.69), secondary(0.42), superior□(0.16)}] ^
[Searching □ {search_no(1.00), search_yes(0.04)}] ^
[Part/Full □ {part time(0.61), full time(0.98)}]

In this class we observe a quite high value for status "widow" and for category "without education", perhaps associated to old people. Also, a high value for "part time worker".

The difference between c18 and c30 lies in the fact that c18 does not comprehend men 45-54 (included in c30), and c30 does not comprehend women 45-64, (included in c18). This fact accounts for the slight differences in the values assigned to some of the professionals and economic activity categories.

CONCLUSION

Galois lattices constitute a very interesting tool to discover relations between objects and variables. In this paper we propose an extension of lattice theory in order to deal directly with probabilistic data. We have defined two kinds of Galois connections on probabilistic data and we have showed how we can obtain Galois lattices on these data. Two applications with real data are presented to illustrate the theoretical results and the proposed algorithm. As a next step it would be interesting to try to take into account the order between categories when it exists.

The main advantage of the proposed method, comparing to other approaches, is that it allows organising probabilistic data in a concept lattice directly, without any prior transformation. Such prior transformations of the data not only often lead to a lost of information, but they also result in an artificial increase of the size of the data table to be processed.

The limitation of the method lays in the size of the resulting lattice. In fact, the number of nodes increases exponentially with the number of objects and variables. The lattice has to be reduced by keeping automatically relevant information. Many authors are interested by this problem, and define lattice properties, but even with these criteria we have a huge quantity of information, difficult to interpret.

REFERENCES

AGRAWAL R., IMIELINSKI T., SWAMI A., "Mining association rules between sets of items in large databases", *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993.

BARBUT M., MONJARDET B. *Ordre et Classification, Algèbre et Combinatoire*, Tomes I et II, Paris, Hachette, 1970.

BIRKHOFF G., *Lattice theory*, American Mathematical Society Colloquium Publications, Vol. XXV, 1st ed., 1940, 3rd ed., 1967.

BOCK H. H., DIDAY E. (eds), *Analysis of Symbolic Data / Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, 2000.

BORDAT J. P., «Calcul pratique du treillis de Galois d'une correspondance», *Mathématiques et Sciences humaines* 96, 1986, p. 31-47.

BOURNAUD I., *Regroupement conceptuel pour l'organisation des connaissances*, Thése de doctorat, Université Paris 6, 1996.

BRITO P., *Analyse de données symboliques. Pyramides d'héritage*, thèse de doctorat, Université Paris IX Dauphine, 1991.

BRITO P., "Order Structure of Symbolic Assertion Objects", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 6, n° 5, 1994, p. 830-835.

BRITO P., "Symbolic Clustering of Probabilistic Data", *Advances in Data Science and Classification*, Rizzi A., Vichi M., Bock H.-H. (eds), Springer-Verlag, 1998, p. 385-390.

BRUZZESE D., IRPINO A., "Galois Lattices of Modal Symbolic Objects", *Advances in Classification and Data Analysis*, Borra S., Rocci R., Vichi M., Schader M. (eds), Springer Verlag, 2001.

BURUSCO A., FUENTES-GONZALES R., "Construction of the L-Fuzzy Concept Lattice", *Fuzzy Sets and Systems* 97(1), 1998, p. 109-114.

DIDAY E., EMILION R., "Treillis de Galois Maximaux et Capacités de Choquet", *Compte-Rendu à l'Académie des Sciences de Paris*, Série I (Analyse Mathématique), tome 325, n° 3, 1997.

DIDAY E., EMILION R., "Maximal and stochastic Galois lattices", *Discrete Applied Mathematics* 127, 2003, p. 271-284.

DUQUENNE V., "Contextual implications between attributes and some representation properties for finite lattices", Ganter B., Wille R. Wolff K. E. (eds), *Beitrage zur Begriffsanalyse*, 1987, Darmstadt.

DUQUENNE V., GUIGUES J.L., "Familles minimales d'implication informatives résultant d'un tableau de données binaires", *Mathématiques et Sciences humaines* 95, 1986, p. 5-18.

GANTER B., WILLE R., *Formal Concept Analysis – Mathematical Foundations*, Berlin, Springer Verlag, 1999.

GIRARD R., RALAMBONDRAINY H., «Recherche de concepts à partir de données arborescentes et imprécises, *Mathématiques, Informatique et Sciences humaines* 147, 1999, p. 87-111.

GODIN R., MINEAU G.W., MISSAOUI R., MILI H., «Méthodes de classification conceptuelle basées sur les treilllis de Galois et applications», *Revue d'intelligence artificielle* 9(2), 1995, p. 105-137.

GUÉNOCHE A., «Construction du treillis de Galois d'une relation binaire», *Mathématiques, Informatique et Sciences humaines* 109, 1990, p. 41-53.

GUÉNOCHE A., «Hiérarchies conceptuelles de données binaires», *Mathématiques, Informatique et Sciences humaines* 121, 1993, p. 23-34.

HERRMANN, C. S., HÖLLDOBLER, S., STROHMAIER, A., "Fuzzy conceptual knowledge processing", *Proceedings of the ACM Symposium on Applied Computing, Philadelphia*, New York, ACM Press, 1996, p. 628-632.

KUZNETSOV S.O., OBIEDKOV S.A., "Comparing performance of algorithms for generating concept lattices", *J. Exp. Theor. Artif. Intell.* 14, 2-3, 2002, p. 189-216.

LECLERC B., "The residuation model for ordinal construction of dissimilarities and other valued objects", Rapport C.AM.S. P.063, Centre d'Analyse et de Mathématique Sociale, Maison des Sciences et de l'Homme, 1990, Paris.

LECLERC B., "The residuation model for the ordinal construction of dissimilarities and other valued objects", in: ed. Bernard Van Cutsem, *Classification and Dissimilarity Analysis*, Lecture notes in Statistics 93, New York, Springer Verlag, 1994, p.149-172.

MEPHU NGUIFO E., «Une nouvelle approche basée sur le treillis de Galois, pour l'apprentissage de concepts», *Mathématiques, Informatique et Sciences humaines* 124, 1993, p. 19-38.

NORRIS E. M., "An algorithm for computing the maximal rectangles in a binary relation", *Revue Roumaine de Mathématiques Pures et Appliquées*, Vol. 23, 4, 1978, p. 243-250.

PASQUIER N., BASTIDE Y., TAOUIL R., LAKHAL L., "Efficient mining of association rules using closed itemset lattices", *Journal of Information Systems*, 24(1), 1999, p. 25-46.

POLAILLON G., *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme*, Thèse de Doctorat, Université Paris IX Dauphine, 1998(a).

POLAILLON G., "Interpretation and reduction of Galois lattices of complex data", *Advances in Data Science and Classification*, Rizzi A., Vichi M., Bock H.-H. (eds), Springer-Verlag, 1998(b), p. 433-440.

POLAILLON G., DIDAY E., "Reduction of symbolic Galois lattices via hierarchies", *Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis* (KESDA'98), Office for Official Publications of the European Communities, Luxembourg, 1999, p. 137-143.

POLLANDT S., *Fuzzy Begriffe: Formale Begriffsanalyse von unscharfen Daten*, Springer, Berlin-Heidelberg, 1997.

RIGUET J., «Relations Binaires, Fermetures, Correspondances de Galois», *Bulletin de la Société Mathématique de France*, 76, 1948.

SEDGEWICK, *Algorithms in C*, Intereditions, 3rd ed., 1999.

STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N., LAKHAL L., "Intelligent structuring and reducing of association rules with formal concept analysis", *Lecture Notes in Computer Science*, 2001.

WILLE R., *Concept lattices and conceptual knowledge systems*, Preprint 1340, Technische Hochschule Darmstadt, 1990.

WILLE R., "Restructuring lattice theory: an approach based on hierarchies of concepts", *Ordered Sets*, D. Reidel (ed.), Dordrecht-Boston, 1982, p. 445-470.

WOLFF K. E., "Concepts in Fuzzy Scaling Theory: order and granularity", *Fuzzy Sets and Systems* 132(1), 2002, p. 63-75.