Directed binary hierarchies and directed ultrametrics

Israël-César Lerman, Pascale Kuntz

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Directed binary hierarchies and directed ultrametrics.

Israël-César LERMAN — Pascale KUNTZ

N° 6815 — version 2

initial version Janvier 2009 — revised version Février 2009

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--6815--FR+ENG

Directed binary hierarchies and directed ultrametrics.

Israël-César LERMAN* , Pascale KUNTZ†

Thème BIO — Systèmes biologiques
Équipes-Projets Symbiose

Rapport de recherche n° 6815 — version 2‡ — initial version Janvier 2009 — revised version Février 2009 — 25 pages

Abstract: Directed binary hierarchies have been introduced in order to give a graphical reduced representation of a family of association rules. This type of structure extends in a very specific way that underlying binary hierarchical classification. In this paper an accurate formalization of this new structure is studied. A binary directed hierarchy is defined as a set of ordered pairs of subsets of the initial individual set satisfying specific conditions. New notion of directed ultrametricity is studied. The main result consists of establishing a bijective correspondence between a directed ultrametric space and a directed binary hierarchy. Moreover, an algorithm is proposed in order to transform a directed ultrametric structure into a graphical representation associated with a directed binary hierarchy.

Key-words: Oriented ascendant hierarchical classification, Ultrametric spaces, Association rules, Graphical representation

* INRIA Rennes - Bretagne Atlantique, and University of Rennes 1, France.Campus de Beaulieu, 35042 Rennes Cédex — lerman@irisa.fr

† Université de Nantes, Laboratoire d'Informatique de Nantes Atlantique, Equipe COD, Site Polytech'Nantes, La Chantrerie, B.P. 50609, 44306 Nantes Cédex 3, France Pascale.Kuntz@polytech.univ-nantes.fr

‡ Les corrections portées - suite à une relecture attentive - sont peu nombreuses, mineures et de pure forme : des espaces blancs oubliés, des parenthèses oubliées, remplacement d'un terme qualificatif par un autre plus approprié, remplacements de crochets par des parenthèses, etc ...

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Hiérachies binaires et Espaces ultramétriques orientés.

Résumé : Les hiérarchies binaires orientées ont été introduites pour fournir une représentation graphique orientée d'une famille de règles implicatives d'association. Une telle structure étend d'une façon très spécifique celle sous jacente aux arbres binaires hiérarchiques de classification. Nous proposons ici une formalisation précise de ce nouveau type de structure. Une hiérarchie binaire orientée est définie comme une famille de couples (ordonnés) de parties de l'ensemble à organiser remplissant des conditions spécifiques. Une nouvelle notion d'ultramétricité binaire orientée est construite. le résultat fondamental consiste en la mise en correspondance bijective entre une structure binaire ultramétrique orientée et une hiérarchie binaire orientée. De plus, un algorithme est proposé pour passer de la structure ultramétrique à celle graphique d'un arbre binaire orienté et valué.

Mots-clés : Classification ascendante hiérarchique orientée, Espaces ultramétriques, Règles d'association, Représentation graphique

1 Introduction

Given a finite set E of entities (objects or descriptive attributes) and a similarity s, or a dissimilarity d, on E, the common objective of a *hierarchical clustering* is to gather in small classes the most similar entities and to separate the most dissimilar ones in different large classes (e.g. [2, 4, 19]). In the vast majority of cases, the measures s and d are symmetrical. However, stimulated by applications in various area (e.g. co-citation analysis, social exchange, psychology) some authors have considered asymmetric measures (e.g. [21, 5, 20, 26, 23]. Yadohisa [25, 23]) has proposed the concept of asymmetrical agglomerative hierarchical clustering to take into account the asymmetry of the relationships in a dendrogram representation.

A specific case of asymmetrical measures is notably found in data mining when considering association rules. Relative to two boolean attributes a and b, an association rule of the form $a \rightarrow b$ means that "when a is $TRUE$, then usually b is also $TRUE$". Association rules differ from logical rules by tolerating few counter-examples; they evaluate an implicative tendency. This notion commonly appears in real-life corpuses where we can observe few individuals for which a is $TRUE$ and b is $FALSE$ without questioning the general trend to have b when we have a. From the seminal works of Agrawal and al. in [1] 1993 and Manilla and al. in 1994 [13], association rules have become one of the major concepts used in data mining. Many measures have been proposed to quantify the strenght of the rule implicative tendency (see Guillet and Hamilton 2007 for a recent state-of-the-art [11]). The vast majority of them are non symmetrical: if Imp measures the implication degree between entity conjunctions, $Imp\,(a, b) \neq Imp\,(b, a)$.

Different clustering algorithms have been proposed for structuring the rule sets (e.g. Toivonen et al. [24], Lent et al. [15]). Most of them build partitions. But, in order to preserve the intrinsical asymmetry of the measures and to discover relationships at different granularity levels, Gras (1996) [6] has proposed a hierarchical model called afterwards "directed hierarchy" by Gras and Kuntz (2005) [8]. Internal nodes of the binary directed hierarchy can be in a sense "rules of rules": e.g. $(a \rightarrow b) \rightarrow (c \rightarrow d)$ whose premisse $(a \rightarrow b)$ and conclusion $(c \rightarrow d)$ can be rules themselves (see figure 1). We refer to Gras et al. (2008) [7] for applications of this model in data mining and didactics.

A first formalization of the concept of binary directed hierarchy has been proposed in a restricted context (where Imp is the implication intensity (Gras and al. 2008)). In this paper, following a work set about by Lerman (2006, 2007) [16, 17], we reexamine this structure in a deeper more accurate and more complete framework. We first define a binary directed hierarchy as a set of ordered pairs of subsets of E which satisfy specific conditions. Then, we propose a directed version of the ultrametricity and show that in a directed ultrametric space the triangles remain isosceles. In these conditions we establish a new bijection theorem between a binary directed hierarchy and a directed ultrametric structure. Due to the orientation, the bijection requires additionnal conditions not present in the classical case.

More precisely, in section 2, the notion of directed binary hierarchy is defined in terms of a set of directed forks. Some of its basic properties are given. Section 3 is devoted to the notion analysis of directed ultrametrics. By defining the strict directed ultrametricity, we obtain a characterization of this new notion in terms of directed isosceles triangles, for which the basis length is strictly lower

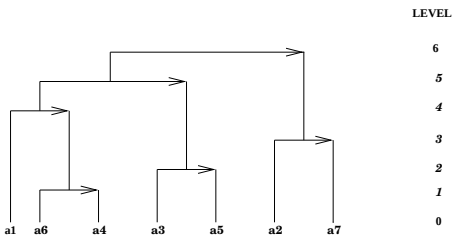

Figure 1: Tree associated with a binary directed hierarchy

than the common length of the two other sides. The main result of this paper is developed in section 4 where we establish the correspondence between a directed binary hierarchy and an associated directed ultrametric space. This result can be viewed as an extension of those obtained in [2, 12, 18] where a hierarchical classification and an ultrametric numerical or ordinal similarity structure are formally associated. Final remarks and comments are considered in the conclusion (section 5).

Let us end this introduction by mentioning that this paper consists of a development of a short summary given in [14].

2 Directed hierarchies

In this section we introduce the notion of complete binary directed hierarchy on a finite set E. The main result is given in Theorem 1. We show that this structure gives rise to a total order on E.

Definition 1 *A directed fork of E is an ordered pair (X, Y) of non empty disjoint subsets of E such that $X \neq E$ and $Y \neq E$. The top of the directed fork (X, Y) is formally represented by $(X \rightarrow Y, X \cup Y)$.*

Figure 2: Directed fork

$X \rightarrow Y$ does mean a directed junction from the subset X to the subset Y. X and Y designate the two components of the directed fork (X, Y). X and Y are called the *left* and the *right* components of the fork. $X \cup Y$ defines the basis of the directed fork. Thus, the nature of a *fork* is *binary*.

Definition 2 *A binary directed hierarchy $\overrightarrow{H}(E)$ is a set of directed forks of E which satisfies the following property: for each unordered pair $\{(X, Y), (Z, T)\}$ of distinct directed forks belonging to $\overrightarrow{H}(E)$, where $card(X \cup Y) \leq card(Z \cup T)$, we have either $(X \cup Y) \cap (Z \cup T) = \emptyset$ or $X \cup Y \subseteq Z$ or $X \cup Y \subseteq T$.*

$(X \cup Y) \cap (Z \cup T) = \emptyset$ will be called *exclusion* between the two directed forks (X, Y) and (Z, T). On the other hand, $X \cup Y \subseteq Z$ (resp. $X \cup Y \subseteq T$) will be called a *left inclusion* (resp., a *right inclusion*) of the fork (X, Y) into the fork (Z, T).

Proposition 1 *For each directed fork $(X, Y) \in \overrightarrow{H}(E)$, there is no other directed fork $(Z, T) \in \overrightarrow{H}(E)$ such that $X \cup Y = Z \cup T$. In particular, $(Y, X) \notin \overrightarrow{H}(E)$.*

Proof: Let us assume that there exist two directed forks (X, Y) and (Z, T) for which $X \cup Y = Z \cup T$. Then, we have clearly:

$$(X \cup Y) \cap (Z \cup T) \neq \emptyset$$

As a consequence we have either

$$X \cup Y \subseteq Z \text{ or } X \cup Y \subseteq T$$

Obviously, none of inclusions can hold. Because, they are equivalent to either

$$Z \cup T \subseteq Z \text{ or } Z \cup T \subseteq T$$

respectively. But, according to the definition of a fork, Z and T are disjoint and non empty. Hence, the previous equation is not valid. On the other hand, since $Y \cup X = X \cup Y$ then we can derive $Y \cup X \notin \overrightarrow{H}(E)$. □

Corollary 1 *A directed fork (X, Y) for which $X \cup Y = E$ is unique.*

This corresponds to a particular case of Proposition 1. □

Proposition 2 *For every element x of E ($x \in E$), there is at most one directed fork for which the singleton subset $\{x\}$ is one of its components.*

Proof: Suppose that there are two distinct directed forks having each $\{x\}$ as one of its components. For each of them two cases have to be considered according to the position of $\{x\}$ in the ordered subset pair defining the directed fork. In fact, it is sufficient to consider the unique case where $\{x\}$ is the *left* component of both directed forks. The proof for the three other cases is strictly analogous.
In these conditions, let us denote by $(\{x\}, Y)$ and $(\{x\}, Z)$ the two distinct directed forks for which $\{x\}$ is their common *left* component. Without loss of generality, let us suppose that $card(Y) \leq card(Z)$. We have

$$(\{x\} \cup Y) \cap (\{x\} \cup Z) \neq \emptyset$$

According to the above Definition 2, we have either

$$\{x\} \cup Y \subseteq \{x\} \text{ or } \{x\} \cup Y \subseteq Z$$

Since $\{x\} \cap Y = \emptyset$, $\{x\} \cap Z = \emptyset$ and $Y \neq \emptyset$, $Z \neq \emptyset$ none of these inclusions can hold. □

There is a natural strict partial order on $\overrightarrow{H}(E)$ that we denote by $\prec_H$: for any pair of directed forks (X, Y) and (Z, T) belonging to $\overrightarrow{H}(E)$, $(X, Y) \prec_H (Z, T)$ if and only if the two following conditions are satisfied:

1. $(X \cup Y) \cap (Z \cup T) \neq \emptyset$
2. $card(X \cup Y) < card(Z \cup T)$

According to Definition 2 one has necessarily :

$$X \cup Y \subset Z \text{ or } X \cup Y \subset T$$

Definition 3 *A binary directed hierarchy is complete when*

1. *for any singleton $x \in E$, there is exactly one directed fork of $\overrightarrow{H}(E)$ such that one component is $\{x\}$;*
2. *there is a directed fork (X, Y) in $\overrightarrow{H}(E)$ such that $X \cup Y = E$;*
3. *for any directed fork (X, Y) in $\overrightarrow{H}(E)$ such that $X \cup Y \neq E$, there exists a directed fork (X', Y') in $\overrightarrow{H}(E)$ such that either $X' = X \cup Y$ or $Y' = X \cup Y$.*

Relative to the above item *2* and from Corollary 1, the directed fork such that $X \cup Y = E$ is necessarily unique. Now, relative to the above item *3* and from Definition 2, the directed fork (X', Y') is necessarily unique. To see this point let us suppose without loss of generality that $X' = X \cup Y$ and assume that there exists another fork (X'', Y'') such that we have either:

$$X'' = X \cup Y \text{ or } Y'' = X \cup Y$$

Without loss of generality one may suppose that $card(X'' \cup Y'') \leq card(X'' \cup Y'')$. In these conditions and from Definition 2, if $X'' = X \cup Y$, then we have necessarily $X'' \cup Y'' \subset X' = X \cup Y$ and this inclusion cannot hold ($(X'' \cap Y'') = \emptyset$). On the other hand, if $Y'' = X \cup Y$, we have necessarily $X'' \cup Y'' \subset Y' = X \cup Y$ and, for the same reason this inclusion cannot hold. □

(X', Y') will be called the *mother* of (X, Y). Finally, according to the strict partial order $\prec_H$ it is easy to establish that there cannot exist a directed fork (Z, T) strictly between (X, Y) and (X', Y'); that is to say such that:

$$(X, Y) \prec_H (Z, T) \prec_H (X', Y')$$

Proposition 3 *Let $\overrightarrow{H}(E)$ be a complete binary directed hierarchy. Given an unordered pair $\{x, y\}$ from E ($\{x, y\} \in P_2(E)$), there exists necessarily a directed fork in $\overrightarrow{H}(E)$ such that one component contains x and the other one contains y.*

Proof: Let us consider x as one of the two elements of $\{x, y\}$. From Definition 3 there is exactly one directed fork of $\overrightarrow{H}(E)$ such that one component is $\{x\}$. Initializing with this directed fork, consider the increasing sequence of forks such that each fork is the *mother* of the preceding one. In this sequence of directed forks consider the first one such that its top defines a basis including x and y. Since the basis of the final directed fork is E, this directed fork does necessarily exist. By construction, each of x and y belongs to one of the two components of the latter directed fork. □

Proposition 4 *If $\overrightarrow{H}(E)$ is a complete binary hierarchy of directed forks, and let $(x, y) \in E \times E$. Assume that there exists a fork (X, Y) for which $(x, y) \in X \times Y$, then this fork is unique.*

Proof: This result can be deduced directly from Definition 2. Assume that we have two different directed forks (X, Y) and (Z, T) such that $(x, y) \in (X, Y)$ and $(x, y) \in (Z, T)$. In these conditions we have:

$$(X \cup Y) \cap (Z \cup T) \neq \emptyset$$

Without loss of generality one may suppose that $card(X \cup Y) \leq card(Z \cup T)$. Then, we necessarily have one of these two alternatives:

1. $X \cup Y \subseteq Z$
2. $X \cup Y \subseteq T$

and neither of them is possible. Indeed, the first (resp., second) contradiction is due to $y \notin Z$ (resp., $x \notin T$). □

Theorem 1 *Let $\overrightarrow{H}(E)$ be a complete binary directed hierarchy on E. The binary relation R_H on E defined by*

$$\forall (x, y) \in E \times E,\ x \neq y,\ xR_Hy \Leftrightarrow \exists! (X, Y) \in \overrightarrow{H}(E), (x, y) \in X \times Y$$

defines a strict total order on E.

Proof:

1. *Antisymmetry* :
Let us consider an arbitrary ordered pair (x, y) belonging to $E \times E$. The conjunction

$$xR_Hy \text{ and } yR_Hx$$

is contradictory. Indeed,

$$xR_Hy \Leftrightarrow \exists!(X, Y) \in \overrightarrow{H}(E) \text{ such that } (x, y) \in X \times Y$$

$$yR_Hx \Leftrightarrow \exists!(Y', X') \in \overrightarrow{H}(E) \text{ such that } (y, x) \in Y' \times X'$$

Clearly,

$$(X \cup Y) \cap (Y' \cup X') \neq \emptyset$$

According to Definition 2 we have at least, one of the following alternatives:

1. $X \cup Y \subseteq X'$;
2. $X \cup Y \subseteq Y'$;
3. $X' \cup Y' \subseteq X$;
4. $X' \cup Y' \subseteq Y$.

Since $x \notin Y$, $x \notin Y'$, $y \notin X$ and $y \notin X'$, any of the four previous alternatives is possible.

2. *Transitivity*: Let us show that

$$\forall (x, y, t) \in E^3, xR_H y \text{ and } yR_H t \Rightarrow xR_H t$$

From the definition of R_H, there are two directed forks (X, Y) and (Y', T) in $\overrightarrow{H}(E)$ such that:

$$(x, y) \in X \times Y \text{ and } (y, t) \in Y' \times T.$$

Since $y \in Y \cap Y'$, $(X \cup Y) \cap (Y' \cup T) \neq \emptyset$. Consequently, from Definition 2, we have either

$$X \cup Y \subset Y' \text{ or } Y' \cup T \subset Y.$$

In the first case $(x, t) \in Y' \times T$ and then $xR_H t$. The same conclusion holds for the second case where $(x, t) \in X \times Y$.

3. *Total order*:
This property is a direct consequence of the Propositions 3 and 4. Let $(x, y) \in E \times E$ with $x \neq y$. By Proposition 3, there exists $(X, Y) \in \overrightarrow{H}(E)$ such that either $(x, y) \in X \times Y$ or $(y, x) \in X \times Y$. If $(x, y) \in X \times Y$ Then by Proposition 4 (X, Y) is unique and so $xR_H y$. Similarly, if $(y, x) \in X \times Y$, Proposition 4 implies $yR_H x$. □

Since R_H defines a total order on E, there exists a unique bijection $\{1, 2, ..., n\} \longrightarrow E$, $i \mapsto x_i$, compatible with the total order (recall $n = card(E)$). Let us denote by I_E the interval on E defined by the totally ordered sequence $(x_1, x_2, ..., x_i, ..., x_n)$.

Corollary 2 *Let (X, Y) be a given directed fork of the complete binary directed hierarchy $\overrightarrow{H}(E)$. By considering the restriction of R_H on X and Y, (X, Y) determines an ordered pair of two consecutive subintervals of $(x_1, x_2, ..., x_i, ..., x_n)$.*

Proof: A recursion argument allows to establish this property that we can call *interval split property.*

Consider the top fork that we denote (X_1, Y_1) whose basis is $X_1 \cup Y_1 = E$. By definition:

$$\forall (x_1, y_1) \in X_1 \times Y_1 \ , \ x_1 R_H y_1$$

By denoting I_{X_1} and I_{Y_1} the subintervals defined by the restrictions of R_H on X_1 and Y_1 respectively, I_{X_1} and I_{Y_1} are necessarily two connex, disjoint and complementary intervals of I_E. I_{X_1} (resp., I_{Y_1}) is a begining (resp., ending) subinterval of I_E. If I_{X_1} has the form $(x_1, x_2, ..., x_c)$, then I_{Y_1} has the form $(x_{c+1}, x_{c+2}, ..., x_n)$. $X_1 \cup Y_1 = E$ and $X_1 \cap Y_1 = \emptyset$.

Now, let us consider an arbitrary fork (X, Y) belonging to $\overrightarrow{H}(E)$. There exists a unique sequence of ordered forks:

$$\big((X_1, Y_1), (X_2, Y_2), ..., (X_m, Y_m)\big)$$

beginning with the initial fork (X_1, Y_1) and ending with the fork $(X_m, Y_m) = (X, Y)$ such that (X_j, Y_j) is the mother fork of (X_{j+1}, Y_{j+1}), $1 \leq j \leq m-1$.

Assume that the *interval split property* is established till (X_m, Y_m) and consider the two forks associated with X_m and Y_m, respectively. Now, let us denote by (X'_{m+1}, X''_{m+1}) the fork whose basis is X_m and by (Y'_{m+1}, Y''_{m+1}) that whose basis is Y_m. By the same argument used for the fork (X_1, Y_1) derived from E, the restrictions of R_H on X'_{m+1} and X''_{m+1} (resp., on Y'_{m+1} and Y''_{m+1}) determine two connex, disjoint and complementary intervals of I_{X_m} (resp., I_{Y_m}). In these conditions the recursion is established. □

3 Directed ultrametrics

In this section we define the notion of a directed ultrametric space on a finite set E. This notion depends on a notion of dissimilarity on E compatible with a total order R on E. Under these conditions, for $x, y, z \in E$ a directed triangle (x, y, z) is such that xRy and yRz. The main result showed in theorems 2 and 3 consists in establishing that for a directed ultrametric space every directed triangle is isosceles having its basis strictly smaller than its equal sides.

Definition 4 *Let R be a total order on E. A* directed dissimilarity d_R *compatible with R is a mapping $E \times E \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$, where $\mathbb{R}_+$ is the set of non negative numbers, of $E \times E$ on the positive real numbers $\overline{\mathbb{R}}_+$ which satisfies the four following conditions :*

1. *for any $x \in E$, $d_R(x, x) = 0$;*
2. *for any $(x, y) \in E \times E$, $x \neq y$, $0 < d_R(x, y) < +\infty$ if xRy;*
3. *for any $((x, y), (z, t)) \in (E \times E) \times (E \times E)$, $x \neq y$ and $z \neq t$, $d_R(x, y) < d_R(t, z)$ if xRy and zRt;*
4. *for any triple $(x, y, z) \in E^3$ $x \neq y, x \neq z, y \neq z$, such that xRy and yRz, then $d_R(x, z) > min\{d_R(x, y), d_R(y, z)\}$.*

Definition 5 *Let us consider a total order R on E and a compatible directed dissimilarity d_R. Then, d_R is called a directed ultrametric if, for any triple $(x, y, z) \in E^3$ such that xRy and yRz, the three conditions are satisfied:*

1. $d_R(x, y) \leq Max\{d_R(x, z), d_R(y, z)\}$;
2. $d_R(x, z) \leq Max\{d_R(x, y), d_R(y, z)\}$;
3. $d_R(y, z) \leq Max\{d_R(x, y), d_R(x, z)\}$.

Moreover, the directed ultrametric is strict if 1 or 3 is strict.

Definition 6 *Let us consider a total order R on E. A triple $(x, y, z) \in E^3$, $x \neq y \neq z$, forms a directed triangle if xRy and yRz.*

In order to distinguish this specific and important case an R-compatible directed ultrametric will be denoted d_{uR}.

Theorem 2 *Let us consider a total order R on E and a compatible directed ultrametric d_{uR} compatible with R. We have either: $d_{uR}(x, y) < d_{uR}(x, z) = d_{uR}(y, z)$ or $d_{uR}(y, z) < d_{uR}(x, y) = d_{uR}(x, z)$*

Thus, each directed triangle from R is isosceles in the metric space (E, d_{uR}) with basis length strictly smaller than the length of the two equal sides; the basis being either xy or yz.

Proof: Two cases have to be distinguished:

1. $d_{uR}(x, y) \leq d_{uR}(y, z)$
2. $d_{uR}(y, z) \leq d_{uR}(x, y)$

Let us consider the first case where $d_{uR}(x, y) \leq d_{uR}(y, z)$. According to the condition *4* of Definition 4, necessarily,

$$d_{uR}(x, z) > d_{uR}(x, y)$$

Then, *3* of Definition 5 gives:

$$d_{uR}(y, z) \leq d_{uR}(x, z).$$

On the other hand, *2* of Definition 5 gives:

$$d_{uR}(x, z) \leq d_{uR}(y, z).$$

Finally,

$$d_{uR}(y, z) = d_{uR}(x, z) > d_{uR}(x, y).$$

The proof for the second case where $d_{uR}(y, z) \leq d_{uR}(x, y)$ is analogous to that for the first case. The condition *4* of Definition 4 entails:

$$d_{uR}(x,z) > d_{uR}(y,z)$$

Then, *1* of Definition 5 gives:

$$d_{uR}(x,y) \leq d_{uR}(x,z).$$

On the other hand, *2* of Definition 5 gives:

$$d_{uR}(x,z) \leq d_{uR}(x,y).$$

Finally,

$$d_{uR}(x,y) = d_{uR}(x,z) > d_{uR}(y,z).$$

Now, let us notice that for the first case the inequality *1* of Definition 5 is strict, but *3* of Definition 5 is reduced to an equality. On the contrary, for the second case, the inequality *3* of Definition 5 is strict, but *1* of Definition 5 is reduced to an equality.

A reciprocal property of the above theorem can be stated as follows:

Theorem 3 *If R is a total order on E and d_R is a directed dissimilarity such that, each directed triangle (x,y,z) is isosceles having its basis length strictly smaller than the common length of the other sides, then, d_R is strictly ultrametric.*

The *proof* is obvious. □

4 Directed binary hierarchy and directed ultrametricity

Now and of course, the objective consists in obtaining as in the classical case [2, 12, 18] a "bijection theorem" between the new concepts of valuated binary directed hierarchy and strictly directed ultrametric distance. However, due to the order on E introduced by the orientation, the bijection theorem requires additional restrictive conditions.

4.1 Directed ultrametricity associated with a directed binary hierarchy

A family of directed ultrametric distances is associated with a directed hierarchy $\overrightarrow{H}(E)$. Each member of this family is defined by a strictly monotone numerical valuation on $\overrightarrow{H}(E)$ endowed with $\prec_H$. This valuation gives rise to the directed ultrametric distance.

Theorem 4 *Let $\overrightarrow{H}(E)$ be a complete binary directed hierarchy on E. From $\overrightarrow{H}(E)$ can be derived a directed dissimilarity d_{R_H} which satisfies the following property: $\forall(x,y,z), x \neq y \neq z$, for which xR_Hy and yR_Hz , exactly one of the two following conditions holds:*

1. $d_{R_H}(x,y) < d_{R_H}(x,z) = d_{R_H}(y,z)$
2. $d_{R_H}(y,z) < d_{R_H}(x,y) = d_{R_H}(x,z)$

In particular, d_{R_H} is ultrametric and compatible with R_H.

Proof: Let us begin by recalling an elementary property concerning the existence of a strictly monotone valuation of a strict partial order on a finite set. By denoting this set by F and by designating $\prec_F$ the strict partial order on F, there exists an infinity of strictly monotone numerical valuations on F endowed with $\prec_F$.

In these conditions, a numerical positive function ν can be defined on $\overrightarrow{H}(E)$ endowed with the strict partial order $\prec_H$, defined in Section 2:

$$\nu : \overrightarrow{H}(E) \rightarrow \mathbb{R}_+$$

such that,

$$\forall((X,Y),(Z,T)) \in \overrightarrow{H}(E) \times \overrightarrow{H}(E), (X,Y) \prec_H (Z,T) \Rightarrow \nu(X,Y) < \nu(Z,T)$$

Now, let us consider an ordered pair $(x,y) \in E \times E$, $x \neq y$, such that xR_Hy. From the previous Propositions 3 and 4 there is exactly one directed fork $(X,Y) \in \overrightarrow{H}(E)$, such that $(x,y) \in X \times Y$. In these conditions, we set $d_{R_H}(x,y) = \nu(X,Y)$. Otherwise, according to Definition 4, $d_{R_H}(x,x) = 0$ for every $x \in E$ and - for simplicity reasons - we can set $d_{R_H}(y,x) = +\infty$ for every $(x,y) \in E \times E$ such that xR_Hy and $x \neq y$. Since R_H is a total order on E, this defines the map $d_{R_H} : E \times E \rightarrow \overline{\mathbb{R}}_+$ uniquely.

Let us now consider a directed triangle $(x,y,z) \in E^3$ associated with R_H. By definition we have xR_Hy and yR_Hz. Let us denote by (X,Y) and (Y',Z) the two unique directed forks such that $(x,y) \in (X,Y)$ and $(y,z) \in (Y',Z)$, respectively. As $(X \cup Y) \cap (Y' \cup Z) \neq \emptyset$ (this intersection contains at least the element y), two cases are possible:

1. $X \cup Y \subset Y'$
2. $Y' \cup Z \subset Y$

Indeed, Definition2 implies that we have one of the four possibilities: (a) $X \cup Y \subset Y'$, (b) $X \cup Y \subset Z$, (c) $Y' \cup Z \subset X$ and (d) $Y' \cup Z \subset Y$. Case (b) cannot occur since $y \in Y$ and $y \notin Z$ as $Y' \cap Z = \emptyset$, similarly (c) cannot happen as $y \in Y'$ and $y \notin X$.

In case *1*, $(X,Y) \prec_{R_H} (Y,Z)$, thus $\nu(X,Y) < \nu(Y',Z)$, consequently:

$$d_{R_H}(x,y) < d_{R_H}(y,z)$$

Besides, $(x,z) \in Y' \times Z$ and $(y,z) \in Y' \times Z$, then $d_{R_H}(x,z) = d_{R_H}(y,z) = \nu(Y',Z)$. Finally, we have:

$$d_{R_H}(x,y) < d_{R_H}(x,z) = d_{R_H}(y,z)$$

The second case is completely similar. In that case $(Y',Z) \prec_{R_H} (X,Y)$ then $\nu(Y',Z) < \nu(X,Y)$, consequently:

$$d_{R_H}(y,z) < d_{R_H}(x,y)$$

Besides, $(x,y) \in X \times Y$ and $(x,z) \in X \times Y$, then $d_{R_H}(x,y) = d_{R_H}(x,z) = \nu(X,Y)$. Finally, we have:

$$d_{R_H}(y,z) < d_{R_H}(x,y) = d_{R_H}(x,z).$$

□

4.2 Directed binary hierarchy associated with a directed ultrametric dissimilarity

Let (E,R) be a totally ordered set and let $d_{uR} : E \times E \to \overline{\mathbb{R}}_+$ be a compatible directed strict ultrametric dissimilarity (see Definitions 4 and 5). In this paragraph we establish a constructive mapping associating with d_{uR} a valuated directed hierarchy. The valuated directed binary hierarchy is built recursively: at each step a family of directed forks ordered by the inclusion relationship (left inclusion or right inclusion) (see Definition 2).

4.2.1 Definitions

Recall that the strict compatible direct ultrametricity is equivalent to the following isosceles property: for any directed triangle $(x,y,z) \in E^3$ associated with R, we have either $d_{uR}(x,y) < d_{uR}(x,z) = d_{uR}(y,z)$ or $d_{uR}(y,z) < d_{uR}(x,y) = d_{uR}(x,z)$ (see Theorems 2 and 3 in Section 3).

Now, let us denote by

$$(x_1, x_2, ..., x_i, ..., x_n)$$

the ordered sequence of the elements of E, according to the total and complete order R. Thus we have:

$$\forall i, 1 \leq i \leq n-1, x_i R x_{i+1}$$

Consider an interval

$$(x_i, x_{i+1}, ..., x_m) \ (m \leq n)$$

of the above sequence and associate with x_i, the sequence of its distances d_{uR} with the subsequent elements in this interval, namely:

$$\{d_{uR}(x_i, x_j) | i \leq j \leq m\}$$

From the compatibility of d_{uR} with respect to the total order R and from its ultrametricity it follows:

$$\forall i, 1 \leq i \leq n-1, \text{ if } i \leq j \leq n-1, \text{ then } d_{uR}(x_i, x_j) \leq d_{uR}(x_i, x_{j+1}).$$

More precisely (Condition 4 in Definition 4),

$$d_{uR}(x_i, x_{j+1}) > min\{d_{uR}(x_i, x_j), d_{uR}(x_j, x_{j+1})\}$$

Then (x_i, x_{j+1}) cannot be the basis of the directed isosceles triangle (x_i, x_j, x_{j+1}). Therefore,

$$d_{uR}(x_i, x_j) \leq d_{uR}(x_i, x_{j+1})$$

and then, the preceding numerical sequence $\{d_{uR}(x_i, x_j) | i \leq j \leq m\}$ is no decreasing.

Among the distance values $\{d_{uR}(x_i, x_j) | i \leq j \leq m\}$ associated with the ordered set defining the sequence $(x_i, x_{i+1}, ..., x_m)$, we assume that there are in all k_i^m distinct values that we retain and denote them by increasing order:

$$\eta_i^m(0) < \eta_i^m(1) < \eta_i^m(2) < ... < \eta_i^m(h) < ... < \eta_i^m(k_i^m)$$

where $\eta_i^m(0) = 0$ and $\eta_i^m(k_i^m)$ is the biggest value. Under these conditions we have:

$$\begin{cases} \{x \in E | d_{uR}(x_i, x) \leq \eta_i^m(0)\} = & \{x_i\} \\ \{x \in E | d_{uR}(x_i, x) \leq \eta_i^m(k_i^m)\} = & \{x_i, x_{i+1}, ..., x_m\} \end{cases}$$

Let us now consider the following increasing sequence of discrete ordered intervals which constitute circles centered at x_i:

$$\left(X_i^m(h)\right)_{0 \leq h \leq k_i^m} = \left(X_i^m(0), X_i^m(1), ..., X_i^m(k_i^m)\right)$$

where

$$X_i^m(h) = \{x_{i'} \in \{x_i, ..., x_m\} | d_{uR}(x_i, x_{i'}) \leq \eta_i^m(h)\}$$

Notice that the left bound of all these intervals is x_i. The sequence of these intervals is increasing with respect to the inclusion relation. The first interval equals (x_i) and the last one equals $(x_i, x_{i+1}, ..., x_m)$, totally ordered by R. We call x_i the *attraction center* of the above series of circles.

Definition 7 *The series of directed forks* $\{(X_i^m(h), X_i^m(h+1) - X_i^m(h)) | 0 \leq h \leq k_i^m - 1\}$ *defines the* fork decomposition *of the totally R-ordered sequence* $(x_i, x_{i+1}, ..., x_m)$*, endowed with the directed ultrametric dissimilarity* d_{uR}*, with respect to the attraction center* x_i*.*

Definition 8 *The* valuated fork decomposition *of the totally-R ordered sequence $(x_i, x_{i+1}, ..., x_m)$, endowed with the directed ultrametric dissimilarity d_{uR}, with respect to the attraction center x_i is defined by the series of the valuated directed forks*

$$\{\big((X_i^m(h), X_i^m(h+1) - X_i^m(h)), \eta_i^m(h+1)\big) | 0 \leq h \leq k_i^m - 1\},$$

where $\eta_i^m(h+1)$ is the common d_{uR} dissimilarity of two elements belonging to $X_i^m(h)$ and $X_i^m(h+1) - X_i^m(h))$ respectively.

4.2.2 First steps of the hierarchical construction

As the hierarchical building process is recursive we here detail the two first steps.

The **first step** consists in repeating the *fork decomposition* of $(x_1, x_2, ..., x_n)$ endowed with the total order R and the directed ultrametric d_{uR}. In this case the attraction center is x_1 and the increasing sequence of the η values ($d_{uR}(x_1, x_i)$ dissimilarities) can be denoted by

$$\eta_1(0) < \eta_1(1) < \eta_1(2) < ... < \eta_1(h) < ... < \eta_1(k_1)$$

Let us consider the increasing sequence - with respect to the inclusion relation - of the discrete intervals which constitute circles centered at x_1:

$$\{X_1(h) | 0 \leq h \leq k_1\}$$

where

$$X_1(h) = \{x_i | d_{uR}(x_1, x_i) \leq \eta_1(h)\}$$

The first interval contains the single element x_1 and the last one the set E. The set difference $X_1(h+1) - X_1(h)$ between two successive intervals is itself a discrete interval. It can also be interpreted as an *annulus* defined by the difference between the two circles $X_1(h+1)$ and $X_1(h)$ centered at x_1 and having the radii $\eta_1(h+1)$ and $\eta_1(h)$, respectively. Each directed fork $(X_1(h), X_1(h+1) - X_1(h))$ is included in the next one $(X_1(h+1), X_1(h+2) - X_1(h+1))$, (*left inclusion*) and the basis of $(X_1(h), X_1(h+1) - X_1(h))$ is the left component of $(X_1(h+1), X_1(h+2) - X_1(h+1))$, $1 \leq h \leq k_1 - 2$.

Hence, considering the partial order $\prec_H$ associated with the fork inclusion, we have:

$$(X_1(h), X_1(h+1) - X_1(h)) \prec_H (X_1(h+1), X_1(h+2) - X_1(h+1))$$

Proposition 5 $\forall h \in [0, k_1]$ $\forall (x, y) \in \big(X_1(h)\big)^2, \forall (z, t) \in \big(X_1(h+1) - X_1(h)\big)^2$
$d_{uR}(x, z) = d_{uR}(x, t) = d_{uR}(y, z) = d_{uR}(y, t) = \eta_1(h+1)$,
$d_{uR}(x, y) < \eta_1(h+1)$ and $d_{uR}(z, t) < \eta_1(h+1)$.

Proof: By construction we have:

$$d_{uR}(x, z) = d_{uR}(x, t) = \eta_1(h+1).$$

Due to the strict isosceles triangle property, it follows:

$$d_{uR}(z,t) < d_{uR}(x,z) = d_{uR}(x,t) = \eta_1(h+1).$$

On the other hand, since, by construction, $d_{uR}(x,y) \leq \eta_1(h)$ and $d_{uR}(x,z) = \eta_1(h+1)$, the strict isosceles triangle property gives:

$$d_{uR}(x,y) < d_{uR}(x,z) = d_{uR}(y,z) = \eta_1(h+1).$$

Now, by considering the strict isosceles triangle (x,y,t), where we already have $d_{uR}(x,y) < \eta_1(h+1)$ and $d_{uR}(x,t) = \eta_1(h+1)$, we deduce that $d_{uR}(y,t) = \eta_1(h+1)$. □

To the directed fork $(X_1(h), X_1(h+1) - X_1(h))$ is assigned the numerical value $\eta_1(h+1)$ which represents the common value of d_{uR} between two respective elements from $X_1(h)$ and $X_1(h+1) - X_1(h)$. This value is strictly smaller than the next one $\eta_1(h+2)$, assigned to the directed fork $(X_1(h+1), X_1(h+2) - X_1(h+1))$, $0 \leq h \leq k_1 - 2$.

The **second step** consists in repeating *fork decomposition* of the discrete interval $X_1(h^1+1) - X_1(h^1$ where h^1 is the smallest value of h for which the right component of $(X_1(h), X_1(h+1) - X_1(h))$, $1 \leq h \leq k_1 - 1$, contains more than one single element. The first element of $X_1(h^1+1) - X_1(h^1)$ denoted by $x_{i(2)}$ is the new attraction center. As in the first step, the numerical value assigned to each new directed fork is equal to the value d_{uR} between any two elements belonging respectively to its left and right components, respectively. The new directed forks are ranked on the right, after the previous ones according to their occurence order.

4.2.3 Algorithmic construction

The general process consists in recursively repeating from left to right, the *fork decomposition* of the right component of the first encountered directed fork including more than one element and for which the *fork decomposition* has not yet been applied. The new attraction center is the left element of the right component. Each new directed fork is valuated according to the d_{uR} dissimilarity between its two components. The process goes on while there remains a directed fork with a right component of cardinality greater than one and on which any *fork decomposition* has yet been applied.

All the new built directed forks are valuated respectively according to their occurence order and ranked on the right in order to obtain a global sequence of the already built forks.

Algorithm DH (Directed-Hierarchy)

Let us define a ϕ structure as an ordered pair $\big((X,Y),\eta\big)$ whose second component η is a numerical value and whose first component (X,Y) is itself an ordered pair of two consecutive and disjoint intervals of the sequence $(x_1, x_2, ..., x_n)$. By denoting *v-fork-decomposition* the valuated fork decomposition function defined in the previous Definition 8, one can associate with the

v-fork-decomposition$\big((x_i, x_{i+1}, ..., x_m)\big)$ $(1 \leq i < m \leq n)$ the following **array** variable:

$$\textbf{var } D_i : \textbf{array}[1, k_i^m - 1] \text{ of } \phi$$

Now, let us denote by C the **array** variable which will contain the successive valuated fork decompositions leading to the construction of the directed binary hierarchy. In the latter, there are exactly $n-1$ internal nodes. Consequently, there are exactly $n-1$ fork decompositions and the dimension of C is $n-1$. Thus, we have:

$$\textbf{var } C : \textbf{array}[1, n-1] \text{ of } \phi$$

$$\{\big(C[j] = ((X[j], Y[j]), \eta[j]) \text{ or } C[j] = empty\big)\}$$

{Initialization}

for j:= 1 **to** $n-1$
$C[j]$:=*empty*
endfor
place D_1 in C from the index 1

{Progression and stop rule}

for j:= 1 **to** $n-2$
$Y[j] :=$ second interval component of $C[j]$
if card$(Y[j]) > 1$
$D_j := v - fork - decomposition(Y[j])$
place D_j in C from its first empty cell
endfor

4.2.4 Illustration

Let us denote by $E = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ the set on which a strict total order R and a compatible strict directed ultrametric dissimilarity are given. Assume that R is defined as follows:

$$R : x_1 < x_2 < x_3 < x_4 < x_5 < x_6$$

The compatible strict directed ultrametric dissimilarity is given by the following table. It is easy to check that every directed triangle (there are in all 20) is isosceles with a basis length strictly lower than the common length of both sides:

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	0.9	0.9	0.9	0.9	0.9
x_2	∞	0	0.6	0.7	0.7	0.7
x_3	∞	∞	0	0.7	0.7	0.7
x_4	∞	∞	∞	0	0.3	0.3
x_5	∞	∞	∞	∞	0	0.1
x_6	∞	∞	∞	∞	∞	0

Let us now make explicit the sequence of the different steps of the algorithmic construction $\mathcal{AC}$. With each built directed fork will be associated its valuation ν. The ν value is equal to the common ultrametric dissimilarity between two elements belonging respectively to the two components of the concerned directed fork.

First step: *Valuated fork decomposition* of E, totally ordered, with respect to x_1 as an attraction center: $x_i(1) = x_1$.

The increasing sequence of the η values (d_{uR} dissimilarities) from x_1 becomes here $(0., 0.9)$. The associated increasing sequence of dicrete ordered intervals is:

$$((x_1), (x_1, x_2, x_3, x_4, x_5, x_6))$$

The first sequence of valuated directed forks is:

$$\big(((\{x_1\},\{x_2,x_3,x_4,x_5,x_6\}),\nu=0.9)\big)$$

This sequence is reduced to one directed fork whose valuation is 0.9. Its second component includes more than one single element of E. The latter corresponds to the subinterval (x_2,x_3,x_4,x_5,x_6).

Second step: *Valuated fork decomposition* of (x_2,x_3,x_4,x_5,x_6) with respect to its first element x_2 as an attraction center: $x_i(2)=x_2$.

The increasing sequence of the η values from x_1 is $(0.,0.6,0.7,0.7,0.7)$. The associated increasing sequence of dicrete ordered intervals is:

$$((x_2),(x_2,x_3),(x_2,x_3,x_4,x_5,x_6))$$

This fork decomposition leads to the following sequence of directed forks:

$$\big(((\{x_2\},\{x_3\}),\nu=0.6),((\{x_2,x_3\},\{x_4,x_5,x_6\}),\nu=0.7)\big)$$

The new *global* sequence of built valuated directed forks is:

$$\big(((\{x_1\},\{x_2,x_3,x_4,x_5,x_6\}),\nu=0.9),((\{x_2\},\{x_3\}),\nu=0.6),((\{x_2,x_3\},\{x_4,x_5,x_6\}),\nu=0.7)\big)$$

In this sequence the first directed fork whose second component contains more than one single element and for which *valuated fork decomposition* has not been yet considered is the third one. This second component is defined by the interval (x_4,x_5,x_6). The *valuated fork decomposition* has to be applied on it.

Third step: *Valuated fork decomposition* of (x_4,x_5,x_6) with respect to its first element x_4 as an attraction center: $x_i(3)=x_4$. The increasing sequence of the η values is $(0.,0.3)$. The associated increasing sequence of discrete intervals is:

$$((x_4),(x_4,x_5,x_6))$$

This *Valuated fork decomposition* leads to the following sequence of directed forks which contains only one element:

$$\big(((\{x_4\},\{x_5,x_6\}),\nu=0.3)\big)$$

The new *global* sequence of built valuated forks is:

$$\big(((\{x_1\},\{x_2,x_3,x_4,x_5,x_6\}),\nu=0.9),((\{x_2\},\{x_3\}),\nu=0.6),((\{x_2,x_3\},\{x_4,x_5,x_6\}),\nu=0.7),$$
$$((\{x_4\},\{x_5,x_6\}),\nu=0.3)\big)$$

In this sequence the first directed fork whose second component contains more than one single element and for which *Valuated fork decomposition* has not been yet considered is the fourth one. This second component is defined by the interval (x_5,x_6). The *Valuated fork decomposition* has to be applied on it.

Fourth step: *Valuated fork decomposition* of (x_5,x_6) with respect to its first

element x_4 as an attraction center: $x_i(3) = x_4$. The increasing sequence of the η values is $(0., 0.1)$. The associated increasing sequence of discrete intervals is:

$$((x_5), (x_5, x_6))$$

This *Valuated fork decomposition* leads to the following sequence of directed forks which contains only one element:

$$\big(((\{x_5\}, \{x_6\}), \nu = 0.1)\big)$$

The new *global* sequence of built valuated directed forks is:

$$\big(((\{x_1\}, \{x_2, x_3, x_4, x_5, x_6\}), \nu = 0.9), ((\{x_2\}, \{x_3\}), \nu = 0.6), ((\{x_2, x_3\}, \{x_4, x_5, x_6\}), \nu = 0.7),$$

$$((\{x_4\}, \{x_5, x_6\}), \nu = 0.3), ((\{x_5\}, \{x_6\}), \nu = 0.1)\big)$$

In this sequence of directed forks there does no remain any directed fork whose right member contains more than one single element and for which *Valuated fork decomposition* has not been applied. Then, the algorithmic construction process is ended.

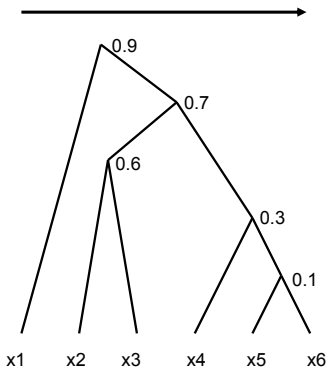

Figure 3: Directed tree associated with the built series of directed forks

4.2.5 Properties

Lemma 1 *At each step of the algorithm DH, the new built directed forks respect either the exclusion or inclusion (either left or right) relationships with the directed forks built at the previous steps.*

Proof: From the recursive nature of the algorithm, it is sufficient to prove this proposition for the first two steps. Four cases can be considerered.

Case 1: (*left inclusion between forks built at step 1*). For the series of directed forks

$$\{(X_1(h), X_1(h+1) - X_1(h)); 0 \leq h \leq k_1 - 1\},$$

the basis of any directed fork constitutes the left component of the next one. Thus, the left inclusion relationship is verified.

Case 2: (*left inclusion between forks built at step 2*). Let us consider the decomposition of $\left(X_1\left(h^1+1\right) - X_1\left(h^1\right)\right)$ considered in the second step of the algorithm, and let us denote by $\left(x_{i(2)}, x_{i(2)+1}, ..., x_{i(2)+k_2}\right)$ the discrete interval which contains $k_2 + 1$elements. Then,

$$d_{uR}\left(x_{i(2)}, x_j\right) \leq d_{uR}\left(x_{i(2)}, x_{j'}\right),$$

for $i(2) < j < j' \leq i(2) + k_2$. Hence, by construction, the left inclusion is checked for all the new built directed forks on $\left(x_{i(2)}, x_{i(2)+1}, ..., x_{i(2)+k_2}\right)$ with the attraction center $x_{i(2)}$.

Case 3: (*right inclusion between forks built at step 2 and forks built at step 1*). The new built directed forks on $\left(x_{i(2)}, x_{i(2)+1}, ..., x_{i(2)+k_2}\right)$ with the attraction center $x_{i(2)}$ respect the right inclusion with the forks obtained from the fork decomposition of $(x_1, x_2, ..., x_n)$, with the attraction center obtained from the fork decomposition of $(x_1, x_2, ..., x_n)$, with the attraction center $x_{i(1)} = x_1$.

Case 4: (*exclusion between forks with a right component of cardinality one*). The directed forks built on X_1^h whose the right component is restricted to a single element are disjoint from the built forks whose basis is included in $\left(X_1\left(h^1+1\right) - X_1\left(h^1\right)\right)$.

Lemma 2 . *Let (x_i, x_j)be any ordered pair of E ($x_i R x_j$). In the directed fork set built by the algorithm DH, there exists a unique directed fork (X, Y)such that $(x_i, x_j) \in X \times Y$.*

Proof: As the initial fork decomposition starts on $(x_1, x_2, ..., x_n)$ which contains $\{x_i, x_j\}$, there exists at least one directed fork buit by the algorithm DH which contains x_i and x_j. Let us denote by $(X_l(g), X_l(g+1) - X_l(g))$ the smallest one : $\{x_i, x_j\} \in X_l(g+1)$ and $card(X_l(g+1))$ is minimum among the set of directed forks built by the DH algorithm.

$X_l(g)$ can not contain a single element which is not x_i. Indeed, in this case, $\{x_i, x_j\} \subset X_l(g+1) - X_l(g)$, which is in contradiction with the assumption

: the fork decomposition of $X_l(g+1) - X_l(g)$ would lead to a directed fork whose basis contains x_i and $x_j$and whose cardinality is strictly smaller than $card(X_l(g+1))$.

$X_l(g)$ can not contain more than one single element such that $\{x_i, x_j\} \subset X_l(g)$. Otherwise, we would have $card(X_l(g)) < card(X_l(g+1))$.

If $X_l(g)$ contains one single element which is x_i, then necessarily x_j belongs to $X_l(g+1) - X_l(g)$.

$X_l(g+1) - X_l(g)$ can not contain one single element which is not x_j. Otherwise, we would have $\{x_i, x_j\} \subset X_l(g)$ and $card(X_l(g)) < card(X_l(g+1))$.

$X_l(g+1) - X_l(g)$ can not contain more than one single element s.t. $\{x_i, x_j\} \subset X_l(g+1) - X_l(g)$. Otherwise, the fork decomposition of $X_l(g+1) - X_l(g)$would lead to a directed fork whose basis contains x_i and $x_j$and whose cardinality is strictly smaller than $card(X_l(g+1))$.

Consequently, we necessarily have $(x_i, x_j) \in X_l(g) \times (X_l(g+1) - X_l(g))$ and $d_{uR}(x_i, x_j)$ is the common value between any pairs of elements from respectively $X_l(g)$ and $X_l(g+1)$.

Theorem 5 *To a total order R on E and a compatible directed ultrametric dissimilarity d_{uR} on E, the algorithmic construction **DH** associates a unique complete directed binary hierarchy valuated by d_{uR}.*

The proof is directly derived from the above Property 5 and Lemmas 1 and 2. □

5 Conclusion and further work

Directed binary hierarchy is a new combinatorial structure devoted to data organization in case where a specific asymmetrical similarity measure is relevant. This structure has been characterized and studied in terms of a set of directed forks. A specific notion of directed ultrametric depending on a total order on the organized set, has been defined and studied with respect to several aspects. An important point concerns its characterization in terms of directed isosceles triangles. The main result obtained here consists of establishing the correspondence between these two previous structures. As mentioned in the introduction, this result is an extension of the "bijection theorem" obtained in the classical symmetrical case [2, 12, 18]. However, for the latter, two versions can be distinguished: *numerical* [2, 12] and *ordinal* [18]. For the first, the correspondence is established up to a strictly increasing function on the ultrametric distances. But, for the second version, the bijection associates two equally finite spaces, the represented set and the representation set. Ordinal directed ultrametric space can also be considered in this work. One other facet may consist in establishing a formal correspondence between the formalization of this new specific hierarchical structure and those considered in the literature. Thus, it is easy to establish that if in the metric space there does not exist equilateral triangle,

then Condition 4 of Definition 4 is more general than Robinson condition [22] adapted to an asymmetrical dissimilarity. On the other hand, one can relate our formal presentation to the classification formalization in terms of a hypergraph $(E, \mathcal{K})$ (see [3]). In our case as in the ascendant hierarchical classification, E is a set of elements and represents the vertice set.The edge set $\mathcal{K}$ is defined by all subsets of E corresponding to the tops of the different forks. Due to the orientation specific results or specific formalization can be obtained in the case of directed hierarchies.

The objective is now to define a complete framework to build and analyse directed hierarchies for real-life applications. It requires aggregation criteria adapted to the directed forks, an efficient algorithm and indicators for interpretation aiding. A first attempt was proposed in [6, 10] in the context of rule mining. Besides, a mathematical and statistical analysis of different criterion types is provided in [16]. However, the aggregation criterion and the indicators defined to study the directed hierarchy are strongly dependent on the implicative statistical analysis framework in which all these works have been developed (see [9]). We believe that the proposed algorithm is general enough to be easily adapted to any kind of dissimilarity.

ACKNOWLEDGEMENT

We are indebted to Professor Benjamin Enriquez of the University Louis Pasteur of Strasbourg for his relevant remarks and suggestions which have improved the presentation of this article.

References

[1] R. AGRAWAL, T. IMIELINSKY, and A. SWAMI. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD'93*, pages 207–216, 1993.

[2] J-P. BENZÉCRI. *L'Analyse des Donnés : La Taxinomie, vol. 1.* Dunod, 1973.

[3] F. BRUCKER and J-P. BARTHÉLEMY. *Eléments de classification.* Hermes, 2007.

[4] A.D. GORDON. *Classification.* Chapmann & Hall, 1999.

[5] R. GRAS. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Doctorat d'État.* PhD thesis, Université de Rennes 1, 1979.

[6] R. GRAS, S. AG ALMOULOUD, M. BAILLEUL, A. A. LARHER, M.and RATSIMBA-RAJOHN H. POLO, and A. TOTOHASINA. *L'implicationstatistique - Nouvelle méthode exploratoire de données.* La Pensée Sauvage, 1996.

[7] R. GRAS, F. GUILLET, F. SPAGNOLO, and E. SUZUKI. *Statistical Implicative Analysis, Studies in Computational Intelligence, 27.* Springer, 2008.

[8] R. GRAS and P. KUNTZ. Discovering r-rules with a directed hierarchy. *Soft Computing*, (10):453–460, 2005.

[9] R. GRAS and P. KUNTZ. An overview of the statistical implicative analysis (sia) development. In F. Spagnolo E. Suzuki R. Gras, F. Guillet, editor, *Statistical Implicative Analysis, Studies in Computational Intelligence*. Springer, 2008.

[10] R. GRAS, P. KUNTZ, and J.-C. RÉGNIER. Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative. In Université de Bordeaux, editor, *11-èmes Rencontres de la Société Francophone de Classification*, 2004.

[11] F. GUILLET and H.J. eds HAMILTON. *Quality measures in data mining, Studies in Computational Intelligence, vol. 43*. Springer, 2007.

[12] S.C. JOHNSON. Hierarchical clustering schemes. *Psychometrika*, (32):241–254, 1967.

[13] M. KLEMENTTINEN, H. MANILLA, P. RONKAINEN, H. TOIVONEN, and A.I. VERKAMO. Finding interesting rules from large sets of discovered association rules. In ACM, editor, *Proceedings of the 3rd International Conference on Information and Knowledge Management*, pages 401–407.

[14] P. KUNTZ and I-C. LERMAN. Directed binary hierarchies and directed ultrametrics. In Université de Naples, editor, *First joint meeting of the Société Francophone de Classification and the Classification and Data Analysis group of the Italian Statistical Society*, pages 337–340, 2008.

[15] B. LENT, A.N. SWAMI, and J. WIDOW. Clustering association rules. In *Proc. of the 13th Conf. on Data Engineering*, pages 220–231.

[16] I-C. LERMAN. Analyse logique, combinatoire et statistique de la construction d'une hiérarchie binaire implicative; niveaux et noeuds significatifs. Publication Interne, a revised version is accepted for publication in Mathématiques et Sciences Humaines 1827, IRISA-INRIA, 2006.

[17] I-C. LERMAN. Sur les différentes expressions formelles d'une hiérarchie binaire symétrique ou implicative. In SupTélécom Paris, editor, *Rencontres de la Société Francophone de Classification*, pages 139–142, 2007.

[18] I.C. LERMAN. *Les bases de la classification automatique*. Gauthier-Villars, 1970.

[19] I.C. LERMAN. *Classification et analyse ordinale des données*. Dunod, 1981.

[20] I.C. LERMAN, R. GRAS, and H. ROSTAM. Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématique et Sciences Humaines*, (74-75):5–35, 5–47, 1981.

[21] J. LOEVINGER. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, (61):1–49, 1947.

[22] W. S. ROBINSON. A model for chronological ordering of archeological deposits. *American antiquity*, (16):295–301, 1951.

[23] A. TAKEUCHI, T. SAITO, and H. YADOHISA. Asymmetric agglomerative hierarchical clustering algorithms and their evaluation. *Journal of Classification*, (24):123–143, 2007.

[24] H. TOIVONEN, M. KLMENTTINEN, P. RONKAIREN, K. HATONEN, and H. MANILA. Pruning and grouping discovered association rules. In *Workshop notes of the ECML Workshop on Statistics, Machine Learning and Knowledge Discovering in Databases*, pages 47–52, 1995.

[25] H. YADOHISA. Formulation of asymmetric agglomerative hierarchical clustering and graphical representation of its results. *Bulletin of the Computational Statistics of Japan*, (15):309–316, 2002.

[26] B. ZIELMAN and W.J. HEISER. Models for asymmetric proximities. *British Journal of Methamatical and Statistical Psychology*, (49):127–146, 1996.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Futurs : Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399