Specifying a Galois k-weak hierarchy of probabilistic objects

Bernard Colin [a]

[a] *Département de Mathématiques, Université de Sherbrooke, Sherbrooke J1K-2R1, Canada*

Jean Diatta [b]

[b] *IREMIA, Université de la Réunion, 15 avenue René Cassin - BP 7151, 97715 Saint-Denis messag cedex 9, France*

Richard Emilion [c,*]

[c] *Département de Mathématiques, Université d'Orléans, France*

Abstract

We present a way to specify a Galois k weak hierarchy of a meet-description context, using, on the one hand, weak clusters associated with a multiway dissimilarity function and, on the other hand, a strict valuation on the entity description space. Moreover, we place ourselves in the framework of a probabilistic meet-closed description context, i.e., a meet-closed description context where entity descriptions are probability distributions. Then we show that, in such a context, the expectation is a strict valuation when the entity description space is endowed with the stochastic order. We also discuss the case of internet traffic flows, as to an illustration of a probabilistic meet-closed description context. Furthermore, we provide examples

of multiway dissimilarity functions on the entity set of a probabilistic meet-closed description context, based on mutual information.

Key words: Cluster, Divergence, Expectation, Galois lattice, Internet flow, Multiway dissimilarity, Mutual information, Probability distribution, Stochastic order, Weak hierarchy.

---

1 Introduction

The data set dealt with in many applications can generally be represented as a meet-closed description context; that is, a context consisting of entities whose descriptions belong to a meet-closed set, the entity description space [1]. Representing data in this way allows to take advantage from both dissimilarity-based data analysis approaches and order-theoretic ones. This paper is concerned with a combination of a dissimilarity-based approach for clustering probabilistic objects, with an order-theoretic one.

We consider the k-weakly hierarchical cluster structure extending the well-known hierarchical one [2], as well as the way it is associated with multiway dissimilarity functions via weak clusters [3, 4]. We also present a way to specify a Galois k-weak hierarchy of a meet-description context, using, on the one hand, weak clusters associated with a multiway dissimilarity function and, on the other hand, a strict valuation on the entity description space. Moreover,

* Corresponding author.
Email addresses: Bernard.Colin@USherbrooke.ca (Bernard Colin),
Jean.Diatta@univ-reunion.fr (Jean Diatta),
richard.emilion@univ-orleans.fr (Richard Emilion).

we place ourselves in the framework of a probabilistic meet-closed description context, i.e., a meet-closed description context where entity descriptions are probability distributions. Then we show that, in such a context, the expectation is a strict valuation when the entity description space is endowed with the stochastic order. We also discuss the case of internet traffic flows , as represented in [5], as to an illustration of a probabilistic meet-closed description context. Furthermore, we provide examples of multiway dissimilarity functions on the entity set of a probabilistic meet-closed description context, based on mutual information. The rest of the paper is organized as follows.

Section 2 is concerned with k-weakly hierarchical cluster structures and the specification of the Galois k-weak hierarchy associated with a k-way dissimilarity function. Probabilistic meet-closed description contexts are addressed in Section 3, where the expectation is shown to be a strict valuation. Finally, examples of multiway dissimilarity functions based on mutual information are discussed in Section 4.

2 Galois k-weak hierarchies and multiway dissimilarities

2.1 k-weak hierarchies

Main cluster structures dealt with in data analysis range from well-known hierarchies to weak hierarchies. Weak hierarchies have been generalized to k-weak hierarchies [3, 4, 2], where k is a positive integer. Weak hierarchies have been introduced by weakening the characteristic condition of so-called strong hierarchies [6].

Let E be a finite nonempty set. A *strong hierarchy* on E is a collection $\mathcal{H}$ of subsets of E, satisfying:

(HI) two members X, Y of $\mathcal{H}$ are always either disjoint or nested, i.e. $X \cap Y \in \{\varnothing, X, Y\}$.

A *weak hierarchy* on E is a collection $\mathcal{H}$ of subsets of E satisfying:

(WH) the intersection of any three members X, Y, Z of $\mathcal{H}$ is always the intersection of two among these three, i.e., $X \cap Y \cap Z \in \{X \cap Y, X \cap Z, Y \cap Z\}$.

In contrast to hierarchies, members of a weak hierarchy can overlap. It should be noticed that every hierarchy is clearly a weak hierarchy. Moreover, condition (WH) generalizes naturally, giving rise to k-weakly hierarchical collections [3, 4], where k is a positive integer. A *k-weak hierarchy* on E is a collection $\mathcal{H}$ of subsets of E satisfying:

(KW) the intersection of any $(k+1)$ members $C_1, \ldots, C_k, C_{k+1}$ of $\mathcal{H}$ is always the intersection of k among these $k+1$, i.e. there is $i \in \{1, \ldots, k+1\}$ such that $\bigcap_{j \neq i} C_j \subseteq C_i$.

It may be noted that a weak hierarchy is nothing else than a 2-weak hierarchy and a 1-weak hierarchy is a chain, i.e., a collection of nested sets.

2.2 The k-weak hierarchy associated with a multiway dissimilarity function

According to the set-based definition given in [1], a *k-way dissimilarity* on E will be any nonnegative real valued and isotone map defined on the set of all nonempty subsets of E with at most k elements, i.e., any map $d : E^*_{\leq k} \to \mathbb{R}_+$

such that $d(X) \leq d(Y)$ when $X \subseteq Y$.

Dissimilarity functions play an important role in cluster analysis where they are often used for constructing clusters having a weak within-cluster and/or a strong between-cluster dissimilarity degrees. Weak clusters introduced in [6] in the framework of pairwise similarity measures are among these clusters. They are said to be weak in contrast to so-called strong clusters.

A subset X of E is said to be a *strong* cluster associated with a pairwise dissimilarity function d_2 (or d_2-*strong* cluster), if its d_2-*strong isolation index*

$$\mathbf{i}^s_{d_2}(X) := \min_{\substack{x,y \in X \\ z \notin X}} \{d_2(x,z) - d_2(x,y)\}$$

is strictly positive. In other words, for all x, y within the cluster and z outside, each of the dissimilarities $d_2(x,z)$ and $d_2(y,z)$ is greater than the dissimilarity $d_2(x,y)$.

A nonempty subset X of E is said to be a *weak* cluster associated with a pairwise dissimilarity function d_2 (or d_2-*weak* cluster), if its d_2-*weak isolation index*

$$\mathbf{i}^w_{d_2}(X) := \min_{\substack{x,y \in X \\ z \notin X}} \{\max\{d_2(x,z), d_2(y,z)\} - d_2(x,y)\}$$

is strictly positive. In other words, for all x, y within the cluster and z outside, at least one of the dissimilarities $d_2(x,z)$ and $d_2(y,z)$ is greater than the dissimilarity $d_2(x,y)$.

It should be noticed that any d_2-strong cluster is a d_2-weak one. Moreover, the notion of weak cluster has been naturally extended to multiway dissimilarity functions [3, 4]. A nonempty subset X of E is said to be a *weak* cluster associated with a k-way dissimilarity measure d_k (or d_k-*weak* cluster) if its

d_k-weak isolation index

$$\mathbf{i}^w_{d_k}(X) := \min_{\substack{Y \in X^*_{\leq k} \\ z \notin X}} \{ \max_{Z \in Y^*_{\leq k-1}} d_k(Z + z) - d_k(Y) \}$$

is strictly positive. On the other hand, it is easily shown that the strong (resp. weak) clusters associated with a pairwise (resp. k-way) dissimilarity function form a strong (resp. k-weak) hierarchy [6, 3, 4].

Proposition 1 *For $k \geq 2$, let d_k be a k-way dissimilarity function on E. Then*

(i) The strong clusters associated with d_2 form a strong hierarchy called the strong hierarchy associated with d_2.

(ii) The weak clusters associated with d_k form a k-weak hierarchy called the k-weak hierarchy associated with d_k.

2.3 Galois k-weak hierarchies

2.3.1 The Galois lattice of a meet-closed description context

A meet-closed description context is a context where entities from a finite set are described in a meet-semilattice. We will denote such a context as a triple $(E, \mathcal{D}, \delta)$, where E is the entity set, $\mathcal{D}$ the entity description space, and δ a descriptor that maps E into $\mathcal{D}$. A meet-closed description context $\mathbb{K} := (E, \mathcal{D}, \delta)$ induces a Galois correspondence between $(\mathcal{P}(E), \subseteq)$ and $\mathcal{D}$ by means of the maps

$$f : X \mapsto \wedge \{\delta(x) \, : \, x \in X\}$$

and

$$g : \omega \mapsto \{x \in E \, : \, \omega \leq \delta(x)\},$$

for $X \subseteq E$ and $\omega \in \mathcal{D}$. The Galois correspondence (f, g) induces, in turn, a closure operator $\varphi_\delta := g \circ f$ on $(\mathcal{P}(E), \subseteq)$.

A subset X of E is said to be *φ_δ-closed* (or a *Galois closed entity set* (of $\mathbb{K}$) under φ_δ) when $\varphi_\delta(X) = X$. Let $G(\mathbb{K})$ denote the set of all pairs $(X, \omega) \in \mathcal{P}(E) \times \mathcal{D}$ such that $\varphi_\delta(X) = X$ and $f(X) = \omega$. Then $G(\mathbb{K})$, endowed with the order defined by $(X_1, \omega_1) \leq (X_2, \omega_2)$ if and only if $X_1 \subseteq X_2$ (or, equivalently $\omega_2 \leq \omega_1$), is a complete lattice called the *Galois lattice* of the context $\mathbb{K}$ [7].

2.3.2 The Galois k-weak hierarchy associated with multiway dissimilarity function

Let $\mathbb{K} := (E, \mathcal{D}, \delta)$ be a meet-closed description context. A *Galois k-weak hierarchy* of $\mathbb{K}$ will be any sub-collection of $G(\mathbb{K})$ consisting of pairs (X, ω), such that the Galois closed entity sets X form a k-weak hierarchy.

A number of Galois k-weak hierarchies can be derived from a given meet-closed description context. In this section, we consider Galois weak hierarchies consisting of pairs (X, ω) such that X is a weak cluster associated with some multiway dissimilarity function.

Let d be a k-way dissimilarity function on E. Recall that the k-weak hierarchy associated with d is the collection of d-weak clusters. We define the *Galois k-weak hierarchy associated with d* to be the maximum sub-collection of $G(\mathbb{K})$ consisting of pairs (X, ω) such that X is a d-weak cluster.

2.3.3 A characterization of Galois closed entity sets

The characterization of Galois closed entity sets given below uses the notion of valuation. A *valuation* on a poset $(P, \leq)$ is a map $h : P \to \mathbb{R}_+$ such that $h(x) \leq h(y)$ when $x \leq y$. A *strict* valuation is a valuation h such that $x < y$ implies $h(x) < h(y)$.

Let $\mathbb{K} := (E, \mathcal{D}, \delta)$ be a meet-closed description context. For any $X \subseteq E$, $\delta(X)$ will denote the set of descriptions of entities belonging to X, and for any $x \in E$, $X + x$ will denote $X \cup \{x\}$. For any positive integer k, let $\mathcal{I}_k(E)$ denote the set of meets of descriptions of nonempty subsets of E with at most k elements, i.e.:

$$\mathcal{I}_k(E) = \{\delta(x_1) \wedge \cdots \wedge \delta(x_k) : x_1, \ldots, x_k \in E\}.$$

It may be noted that $\mathcal{I}_1(E) = \delta(E)$ and $\mathcal{I}(E) := \mathcal{I}_{|E|}(E)$ is the meet-semilattice generated by $\delta(E)$. The result below characterizes the Galois closed entity sets of $\mathbb{K}$, in terms of strict valuations.

Proposition 2 *A subset X of E is φ_δ-closed if and only if for any strict valuation h on $\mathcal{I}_{|E|}(\delta(E))$ and for any $x \in E$: $h(\wedge\delta(X + x)) = h(\wedge\delta(X))$, where $|E|$ denotes the cardinality of E.*

Finally, according to Propositions 1 and 2, a Galois k-weak hierarchy GkW of the meet-closed description context $\mathbb{K}$ can be specified in the following way:

(1) Define a k-way dissimilarity function d on E;
(2) Define a strict valuation h on $\mathcal{D}$;
(3) Set GkW to be the collection of pairs $(X, \wedge\delta(X))$, where X is a φ_δ-closed d-weak cluster.

3 Strict valuations for a probabilistic meet-closed description context

3.1 Stochastic order and lattice structure

We wish now to apply the preceding results to objects having a probabilistic behavior. This will extend the notion of stochastic Galois Lattice introduced in [8] to other cluster systems. The probabilistic objects can be individuals (for example a customer or a group of customers in economy) as well as complex devices (for example queuing systems in a computer network). The behavior of each object is modeled by a random variable (r.v.) but we are mainly focused on the distribution of this r.v.. The following notations agree with the ones used in the preceding sections.

Let E denote a set which elements x are called probabilistic objects. To each x is associated a r.v. V_x taking its values in the real line $\mathbb{R}$ and defined on a probability space $(\Omega, \mathcal{F}, P)$ which may depend on x. The distribution of this r.v. is the probability measure P_x on $\mathbb{R}$ such that $P_x(A) = P(V_x \in A)$ for any Borel subset $A \subseteq \mathbb{R}$.

In the example developed in the next section, E will denote a set of internet flows, x a flow and P_x a probability measure on an interval $[0, B]$ of bandwidth.

The set $\mathcal{D}$ of descriptions of the elements of E is therefore the set $\mathcal{D}(\mathbb{R})$ of all probability measures on $\mathbb{R}$. Let δ denote the mapping

$$\delta : E \to \mathcal{D} \text{ defined as } \delta(x) = P_x.$$

We will write $x \sim y$ (or $V_x \sim V_y$) whenever V_x and V_y have the same distribution,

that is $P_x = P_y$.

We will consider the survival function

$$F_x^*(t) = P_x((t, +\infty)) = P(V_x > t), t \in \mathbb{R}$$

which is equal to $1 - F_x$ where F_x denotes the usual cumulative distribution function (cdf) of V_x defined by

$$F_x(t) = P_x((-\infty, t]) = P(V_x \leq t), t \in \mathbb{R}.$$

As seen previously, we need in an essential way a semi-ordering relation on $\mathcal{D}$. Several definitions are possible (see e.g.[9], [10]) but we will use here the stochastic order, mainly because of the proposition mentioned below.

We will say that P_x is stochastically lower than P_y, shortly $P_x \leq_{st} P_y$, if and only if

$$\int_{\mathbb{R}} f(s) dP_x(s) \leq \int_{\mathbb{R}} f(s) dP_y(s)$$

for any non decreasing $f : \mathbb{R} \to \mathbb{R}$ such that both integrals exist.

Note that this definition does not require that V_x and V_y be defined on the same $(\Omega, \mathcal{F}, P)$.

It can be proved that an equivalent definition is that the above inequality holds for any non decreasing continuous and positive function $f : \mathbb{R} \to \mathbb{R}_+$ [10, pp. 275-276]. Moreover,

$$P_x \leq_{st} P_y \text{ iff } F_x^* \leq F_y^* \text{ or equivalently } F_y \leq F_x$$

([10]Muller-Scarsini04 pp 275-276) so that $\leq_{st}$ is an ordering relation on $\mathcal{D}(\mathbb{R})$.

Its is clear that the ordered set $(\mathcal{D}(\mathbb{R}), \leq_{st})$ is a lattice with

$$F_x^* \wedge_{st} F_y^* = \min(F_x^*, F_y^*) \text{ and } F_x^* \vee_{st} F_y^* = \max(F_x^*, F_y^*).$$

Now we assume that all the above r.v.'s have a finite first moment. Let h be the valuation defined by

$$h : \mathcal{D} \to \mathbb{R} \ , \ h(P_x) = \int_{\mathbb{R}} t dP_x(t) = E_P(V_x)$$

where E_P denotes the expectation w.r.t. the probability measure P on Ω. It can be noticed that h takes its value in $\mathbb{R}$ and not in $\mathbb{R}_+$. However, as it will be seen later, this is not a problem as we will only deal with finite sets E.

Our clustering methods hinges on the observation that the valuation h is strictly increasing on the lattice $(\mathcal{D}(\mathbb{R}), \leq_{st}, \wedge_{st}, \vee_{st})$:

Proposition 3 *$P_x \leq_{st} P_y$ and $P_x \neq P_y$ implies that $h(x) < h(y)$.*

Proof. We propose two different proofs (i) and (ii).

(i) Let

$$F_x^{-1}(u) = \inf\{t : F_x(t) > u\}$$

be the pseudo-inverse of F_x and let U be a uniform r.v. on on $[0, 1]$. Then

$$V = F_x^{-1}(U) \sim V_x$$

since

$$P(V \leq t) = P(U \leq F_x(t)) = F_x(t) = P(V_x \leq t).$$

Since $V \leq W = F_y^{-1}(U)$, the equality $h(x) = h(y)$ would imply $V = W$ contradicting $P_x \neq P_y$, therefore $h(x) < h(y)$.

(ii) Let $V_x^+ = V_x \vee 0$ and $V_x^- = (-V_x) \vee 0$ be the usual positive parts of V_x. Since

$$E_P(V_x^+) = \int_0^{+\infty} P(V_x^+ > t)dt = \int_0^{+\infty} P(V_x > t)dt$$

and

$$E_P(V_x^-) = \int_0^{+\infty} P(V_x^- \geq t)dt = \int_0^{+\infty} P(V_x \leq -t)dt$$

we have

$$E_P(V_x) = \int_0^{+\infty} (P(V_x > t) - P(V_x \leq -t))dt.$$

If $V_x \leq_{st} V_y$, then

$$P(V_x > t) - P(V_x \leq -t) \leq P(V_y > t) - P(V_y \leq -t)$$

and the equality $E_P(V_x) = E_P(V_y)$ would imply

$$P(V_x > t) - P(V_x \leq -t) = P(V_y > t) - P(V_y \leq -t)$$

for almost all $t > 0$. Therefore

$$P(V_x > t) = P(V_y > t) \text{ and } P(V_x > -t) = P(V_y > -t)$$

that is $P(V_x > t) = P(V_y > t)$ for almost all t. Hence we would have $F_x = F_y$ since these functions are right continuous. This is in contradiction with $P_x \neq P_y$ and thus $h(x) < h(y)$.

Observe that all the arguments above also hold for *categorical* r.v.'s V_x taking their values in any *totally ordered finite or countable set*. Indeed it suffices to define $\leq_{st}$ by ordering the survival functions F^* which are well defined.

Of course the relation $\leq_{st}$ can be defined in $\mathcal{D}(\mathbb{R}^n)$ by testing the integrals on non decreasing $f : \mathbb{R}^n \to \mathbb{R}$. Unfortunately $\mathcal{D}(\mathbb{R}^n)$ need not be a lattice for $n \geq 2$ [9] but a strict valuation can be defined as follows :

$$h_n(P_x) = \sum_{i=1}^{n} c_i h(P_x^i)$$

where P_x^i denotes the distribution of V_x^i for any random vector $V_x = (V_x^1, ..., V_x^i, ..., V_x^n)$ having for distribution P_x, the weights c_i being arbitrary strictly positive real numbers.

Proposition 4 *$P_x \leq_{st} P_y$ in $\mathcal{D}(\mathbb{R}^n)$ and $P_x \neq P_y$ implies $h_n(P_x) < h_n(P_y)$*

Proof. By a theorem of Strassen (see e.g. [10, p. 275]) we know that $P_x \leq_{st} P_y$ in $\mathcal{D}(\mathbb{R}^n)$ iff there exists two r.v. V and W having for distribution P_x and P_y respectively, such that $V \leq W$ *a.s.* Since the coordinates of these vectors are such that $V_i \leq W_i$ *a.s.*, we have $h(P_x^i) = E(V_i) \leq E(W_i) = h(P_y^i)$ so that h is increasing. Moreover since $c_i > 0$ for $i = 1, ..., n$, $h_n(P_x) = h_n(P_y)$ would imply $h(P_x^i) = h(P_y^i)$ and $V_i = W_i$ *a.s.*. Then we would have $V = W$ *a.s.* and $P_x = P_y$, contradicting $P_x \neq P_y$. Thus h is a strict valuation.

Note that for $n = 1$, the relation $\leq_{st}$ can be interpretated as follows : let t be any value and let α_t be the chance that $V_x > t$, then the chance of finding $V_y > t$ is greater than α_t. In other words $P_x \leq_{st} P_y$ means that there is a tendancy of finding V_y greater than V_x. The theorem of Strassen gives a precise statement of this interpretation for any n.

Also note that the proof of Strassen theorem used in the proof of the last proposition is given in the above proof (i) for $n = 1$.

It can be observed that the conclusion of the propositions does no more hold when using other usual ordering relations. Moreover, it may be noted that various dissimilarity functions on probabilistic objects can be found in the literature (see, for instance, [11, p. 69] for pairwise dissimilarities). A canonical multiway dissimilarity function for a meet-closed description context can also be found in [1].

3.2 Internet flows and density of bandwidth occupation time

In the Internet network, the information is stored into small units, namely the packets. An Internet flow is a sequence of packets having a common property, for example the same source and destination and the same protocol. We are interested here to highly aggregated flows. As mentionned in [5], the classification of such flows is of interest for various reasons :

- global view of the Internet traffic

- better management by using a separate treatment for each traffic class

- detection of denial of service attacks since the classes are different wether the network is attacked or not.

Our goal here is to study various cluster structures which are different from the partition obtained in [5] by estimating a mixture of Dirichlet distributions.

Measurements performed on the network yield a representation of a flow as a function $t \rightarrow X_t$ where $t \in [0, T]$ is the time variable on an observation window of size T and X_t is the bandwidth of the flow at time t, that is the size (in bytes) transferred per time unit.

It is highly interesting to consider the distribution of the bandwidth inside this window. More precisely, suppose that the bandwidth lies in an interval $[0, B]$, then, to any subinterval $[a, b]$ of $[0, B]$ we can associate the portion of time that the process X_t has spent in $[a, b]$, that is the amount of time that X_t has spent in $[a, b]$ divided by T. In that way, to each partition of $[0, B]$ into intervals, is associated an histogram of bandwidth occupation time.

Note that if we consider the observed function $t \to X_t$ as a path of a stochastic process, then the above histogram is an estimator of the density of the bandwidth occupation time, that is the local time $L(T, .)$ under hypotheses that ensure the existence of this latter.

Next, we derive from the above histograms, descending cumulative histograms that are estimators of the survival function associated to the distribution of the bandwidth. Their comparisons induce the desired stochastic order between flows.

With respect to the notations of the preceding section we then see that E denote the set of observed flows. Fix a partition of $[0, B]$ into intervals. Let $x \in E$ denote any arbitrary flow and let P_x the probability measure on the interval $[0, B]$ induced by the histogram of bandwidth occupation time of the flow x. This simply means that for any $a, b : 0 \leq a \leq b \leq B$, $P_x[a, b] = \int_a^b H_x(t)dt$ where H_x is the stepwise function defined by the heights of the bars of the histogram. Finally V_x denotes any r.v. having for distribution P_x.

4 Multiway dissimilarity functions based on information theory

If, usually, most statistical models well describe the multivariate reality of the data set by the mean of a geometric approach, less of them take into account the stochastic contents of the data in order to describe another feature of the same reality by the means of probabilistic concepts as: entropy, divergence and mutual information.

For instance, in classification theory, we often use models which are based on metric criteria and we rarely consider models which are based on probabilistic ones, like those deriving from information theory. More precisely, if for a given data set, as introduced in the previous section, we have to choose a similarity or dissimilarity measure between two or more elements which are similar to probability measures, frequencies, positives values or contingency tables, we can take into account this probabilistic content by choosing, among others, a measure based on the notion of divergence between two probability measures.

4.1 Generalized divergence

As a form of recall, we remind the following results presented in details in, for instance, Csiszár [12, 13, 14], Ali and Silvey [15], and by Zakai and Ziv [16].

Let $\varphi(t)$ be any convex function from $\mathbb{R}^+ \setminus \{0\}$ to $\mathbb{R}$. In order to solve some

indetermination, we adopt the following usual conventions:

$$\varphi(0) = \lim_{t\to 0^+} \varphi(t)$$

$$0\varphi(\tfrac{0}{0}) = 0$$

$$0\varphi(\tfrac{a}{0}) = \lim_{\delta\to 0^+} \delta\varphi(\tfrac{a}{\delta}) = a \lim_{\delta\to 0^+} \delta\varphi(\tfrac{1}{\delta}) \quad a > 0$$

The following lemma, is due to Csiszár [14]:

Let $(\Omega, \mathcal{F}, \mu)$ be any measured space (we will suppose however that the measure μ is $\sigma-$finite) and let α and β be two non negative measurable functions defined on $(\Omega, \mathcal{F}, \mu)$. Then:

i)

$$\int \mathbb{I}_A \beta \varphi\left(\frac{\alpha}{\beta}\right) d\mu$$

is defined for each $A \in \Omega$ for which α and β are integrable. Moreover if, for such a set A, $\int \mathbb{I}_A \beta d\mu$ is strictly positive and if $\varphi(t)$ is strictly convex at:

$$t_0 = \varphi\left(\frac{\int \mathbb{I}_A \alpha d\mu}{\int \mathbb{I}_A \beta d\mu}\right)$$

we have:

ii)

$$\int \mathbb{I}_A \beta \varphi\left(\frac{\alpha}{\beta}\right) d\mu \geq \left(\int \mathbb{I}_A \beta d\mu\right) \varphi\left(\frac{\int \mathbb{I}_A \alpha d\mu}{\int \mathbb{I}_A \beta d\mu}\right) > -\infty$$

where the equality holds if and only if $\alpha = t_0\beta \quad \mu - a.s.$ on A.

We will assume in the following, that the equalities are considered in the "almost surely" sense.

We consider now two probability measures $\mu_1$and μ_2 defined on $(\Omega, \mathcal{F})$ such that $\mu_i \ll \mu$ for $i = 1, 2$. We define the $\varphi-$divergence, or the generalized

divergence or merely divergence, of μ_1 with respect to μ_2 by:

$$\begin{aligned} I_\varphi(\mu_1, \mu_2) &= \int \varphi\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_2 \\ &= \int \varphi\left(\frac{f_1}{f_2}\right) f_2 d\mu \quad \text{where } f_i = \frac{d\mu_i}{d\mu} \quad \text{for } i = 1, 2 \end{aligned}$$

The above lemma ensures the existence of $I_\varphi(\mu_1, \mu_2)$ and shows that $I_\varphi(\mu_1, \mu_2) \geq \varphi(1)$, where equality holds if and only if $\mu_1 = \mu_2$ in so far as φ is strictly convex at $t_0 = 1$. Obviously, $I_\varphi(\mu_1, \mu_2)$ does not depend on the choice of μ. Moreover one can, for homogeneous models, write this expression as :

$$\begin{aligned} I_\varphi(\mu_1, \mu_2) &= \int \frac{d\mu_2}{d\mu_1} \varphi\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_1 \\ &= \int \frac{f_2}{f_1} \varphi\left(\frac{f_1}{f_2}\right) f_1 d\mu \end{aligned}$$

The following table presents, for various φ, the main measures of $\varphi-$divergence:

$\varphi(x)$	$Name$
$xLogx$ $-Logx$ $(x-1)\,Logx$	Kullback and Leibler
$\lvert x-1\rvert$	Distance in variation
$x^{\alpha}sgn\,(\alpha-1)$ avec: $0<\alpha\neq 1$	Rényi's divergence of order α
$(\sqrt{x}-1)^2$	Hellinger
$1-x^{\alpha}\ \ 0<\alpha<1$	Chernoff
$(x-1)^2$	χ^2
$\left\lvert 1-x^{1/m}\right\rvert^m\ \ m>0$	Jeffreys
$1-\min(x,1)$	Wald

(For more details see, for instance, Goel [17], Adhikari and Joshi[18], Aczél and Daróczy [19], as well as Rényi[20]). Note that $\varphi(1)=0$ except for Rényi's divergence of order α.

Moreover, the generalized divergence leads to the following concept of mutual information between n random variables (or random vectors).

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, and let $X_1,X_2,...,X_k$ be k random variables defined on $(\Omega,\mathcal{F},\mathbb{P})$ taking values in measured spaces $(\mathcal{X}_i,\mathcal{F}_i,\lambda_i)\ \ i=$

$1, 2, ..., k$. Note by $\mu_{X_1,X_2,...,X_k}$ and by $\otimes_{i=1}^{i=k}\mu_{X_i}$ the probability measures defined on the product space $\left(\times_{i=1}^{i=k}\mathcal{X}_i, \otimes_{i=1}^{i=k}\mathcal{F}_i, \otimes_{i=1}^{i=k}\lambda_i\right)$, respectively equal to the joint probability measure and to the product of the marginal probability measures associated with $X_1, X_2, ..., X_k$ and which are supposed to be absolutely continuous with respect to the product measure $\lambda = \otimes_{i=1}^{i=k}\lambda_i$. We then define the φ-mutual information or the mutual information between the random variables $X_1, X_2, ..., X_k$, by:

$$\begin{aligned}\mathcal{I}_\varphi(X_1, X_2, ..., X_k) &= I_\varphi\left(\mu_{X_1,X_2,...,X_k}, \bigotimes_{i=1}^{k}\mu_{X_i}\right)\\ &= \int \varphi\left(\frac{d\mu_{X_1,X_2,...,X_k}}{d\left(\otimes_{i=1}^{k}\mu_{X_i}\right)}\right) d\left(\bigotimes_{i=1}^{k}\mu_{X_i}\right)\\ &= \int \varphi\left(\frac{f_1}{f_2}\right) f_2 d\lambda\end{aligned}$$

where, f_1 and f_2 are, respectively, the probability density functions of the probability measures $\mu_{X_1,X_2,...,X_k}$ and $\otimes_{i=1}^{i=k}\mu_{X_i}$ with respect to the product measure $\lambda = \otimes_{i=1}^{i=k}\lambda_i$. In most of the cases, the space $\mathcal{X}_i$ is for, every i, either the real line $\mathbb{R}$ endowed with the *Lebesgue* measure, or a discrete space endowed with the counting measure. Some properties of the mutual information follow (for more details and proofs, one can see, among others, Pinsker [21], Mc Eliece [22], Csiszár [12, 14], Gavurin [23]).

i) If $\varphi(1) \geq 0$, then $\mathcal{I}_\varphi(X_1, X_2, ..., X_k) \geq 0$ where the equality holds if and only if the random variables $X_1, X_2, ..., X_k$ are independent.

ii) $\mathcal{I}_\varphi(X_1, X_2, ..., X_k) \geq \mathcal{I}_\varphi(X_1, X_2, ..., X_{k-1})$ where the equality holds if and only if the random variable X_k is independent of the random variables $X_1, X_2, ..., X_{k-1}$.

iii) $\mathcal{I}_\varphi(X_1, X_2, ..., X_k)$ is convex with respect to $\mu_{X_1,X_2,...,X_k}$.

iv) If for every $j = 1, 2, ..., k$, the functions g_j from $\left(\times_{i=1}^{i=k} \mathcal{X}_i, \otimes_{i=1}^{i=k} \mathcal{F}_i\right)$ to $(\mathcal{Y}_j, \mathcal{G}_j)$ are measurable, then we have:

$$\mathcal{I}_\varphi(Y_1, Y_2, ..., Y_k) \leq \mathcal{I}_\varphi(X_1, X_2, ..., X_k)$$

where $Y_j = g_j(X_1, X_2, ..., X_k)$.

This last result, known as the "data-processing theorem", shows that some transformation of the initial variables leads, in general, to a mutual information loss.

v) $\mathcal{I}_\varphi((X_1, X_2, ..., X_{k-1}), X_k) = \mathbb{E}^{\mu_{X_k}}\left(\mathcal{I}_\varphi\left(\mu_{X_1,X_2,...,X_{k-1}|X_k}, \mu_{X_1,X_2,...,X_{k-1}}\right)\right)$ where $\mu_{X_1,X_2,...,X_{k-1}|X_k}$ is the conditional probability measure of the random variables $X_1, X_2, ..., X_{k-1}$ given X_k and where $\mathbb{E}^{\mu_{X_k}}$ denotes the expectation with respect to the probability measure μ_{X_k}.

vi) $\mathcal{I}_\varphi(X_1, X_2, ..., X_{k-1}, X_k) = \mathbb{E}^{\mu_{X_k}}\left(\mathcal{I}_\varphi\left(\mu_{X_1,X_2,...,X_{k-1}|X_k}, \otimes_{i=1}^{i=k-1} \mu_{X_k}\right)\right)$.

4.2 Dissimilarity measure based on divergence

Consider, as it is the case in many applications, a finite family of probability measures definied on the same finite measurable space $(I,\ \mathcal{P}(I))$ with $Card(I)$, equals to r. Using usual modifications, the denumerable and continuous cases are straightforward. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures defined on $\mathcal{P}(I)$. Given a convex function φ from $\mathbb{R}^+ \setminus \{0\}$ to $\mathbb{R}$ fulfilling the conditions stated just above, ones recalls that the generalized divergence (or $\varphi - divergence$) of $\mathbb{P}$ with respect to $\mathbb{Q}$, is given by:

$$D_\varphi(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{r} \varphi\left(\frac{p_i}{q_i}\right) q_i = \mathbb{E}^{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$$

Furthermore $D_\varphi(\mathbb{P},\mathbb{Q}) \geq \varphi(1)$ where the equality holds if and only if $\mathbb{P}=\mathbb{Q}$ if φ is strictly convex for $t=1$.

It is clear that in general, $D_\varphi(\mathbb{P},\mathbb{Q}) \neq D_\varphi(\mathbb{Q},\mathbb{P})$. However, if $\widetilde{\varphi}$ is given by:

$$\widetilde{\varphi}(t) = \varphi(t) + t\varphi\left(\frac{1}{t}\right)$$

Then :

$$\begin{aligned} D_{\widetilde{\varphi}}(\mathbb{P},\mathbb{Q}) &= \sum_{i=1}^{r}\left[\varphi\left(\frac{p_i}{q_i}\right) + \frac{p_i}{q_i}\varphi\left(\frac{q_i}{p_i}\right)\right] q_i \\ &= \sum_{i=1}^{r}\varphi\left(\frac{p_i}{q_i}\right) q_i + \sum_{i=1}^{r}\varphi\left(\frac{q_i}{p_i}\right) p_i \\ &= \mathbb{E}^{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] + \mathbb{E}^{\mathbb{P}}\left[\varphi\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \end{aligned}$$

If for normalization or comparison reasons, one wishes to preserve the inequality $D_{\widetilde{\varphi}}(\mathbb{P},\mathbb{Q}) \geq \widetilde{\varphi}(1)$, one has to take $\widetilde{\varphi}$ such as :

$$\widetilde{\varphi}(t) = \lambda\varphi(t) + (1-\lambda)\, t\varphi\left(\frac{1}{t}\right)$$

where $\lambda \in]0,1[$. It is easy to verify that $D_{\widetilde{\varphi}}(\mathbb{P},\mathbb{Q}) = D_{\widetilde{\varphi}}(\mathbb{Q},\mathbb{P})$ but $D_{\widetilde{\varphi},\lambda}(\mathbb{P},\mathbb{Q}) = D_{\widetilde{\varphi},\lambda}(\mathbb{Q},\mathbb{P})$ if and only if $\lambda = \frac{1}{2}$ which generalizes the notion of information radius as introduced by Sibson (see Sibson [24] as Jardine and Sibson [25]).

4.2.1 Binary dissimilarity measure

In order to measure the dissimilarity between two probability measures, one can consider, as above, the divergence between them. However, the fact that the divergence is not symmetric, and the fact that the symmetrical version of the divergence is not very easy to use, lead us to the following definitions.

Let us consider two probability measures $\mathbb{P}=\{p_j\},\ \mathbb{Q}=\{q_j\}\ j = 1,2,...,p$

absolutely continuous with respect to another probability measure $\mathbb{R} = \{r_j\}$ ($\mathbb{R}$ is a reference measure as for instance an uniform or a counting measure). One then defines the dissimilarity $\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ with respect to $\mathbb{R}$ by:

$$\begin{aligned}\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) &= |D_\varphi(\mathbb{P},\mathbb{R}) - D_\varphi(\mathbb{Q},\mathbb{R})| = \left|\mathbb{E}^{\mathbb{R}}\left(\varphi\left(\frac{d\mathbb{P}}{d\mathbb{R}}\right) - \varphi\left(\frac{d\mathbb{Q}}{d\mathbb{R}}\right)\right)\right| \\ &= \left|\sum_j \varphi\left(\frac{p_j}{r_j}\right) r_j - \sum_j \varphi\left(\frac{q_j}{r_j}\right) r_j\right| \\ &= \left|\sum_j \left(\varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right)\right) r_j\right|\end{aligned}$$

Therefore one has:

$$\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = \Delta_{\varphi,\mathbb{R}}(\mathbb{Q},\mathbb{P})$$

In the finite case it is not even necessary for $\mathbb{R}$ to be a probability measure. For instance, if $\mathbb{R}$ is a counting measure, as it is often the case, then:

$$D_\varphi(\mathbb{P},\mathbb{R}) = \sum_j \varphi(p_j) \ ; \ D_\varphi(\mathbb{Q},\mathbb{R}) = \sum_j \varphi(q_j)$$

which, for $\varphi(t) = tLogt$ gives:

$$\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = \left|\sum_j p_j Log(p_j) - \sum_j q_j Log(q_j)\right|$$

or, if X and Y are some random variables with $\mathbb{P}$ and $\mathbb{Q}$ as probability measures,

$$\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = \Delta_{\varphi,\mathbb{R}}(X,Y) = |H(X) - H(Y)|$$

where $H(X)$ and $H(Y)$ are respectively the entropies of X and Y. Moreover if $\mathbb{R}$ is any probability measure, then for $\varphi(t) = tLogt$ one has:

$$\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = \left|H(X) - H(Y) + \sum_j (p_j - q_j) Log(r_j)\right|$$

Also, if $\mathbb{M} = \{m_i\}$ is another probability measure (absolutely continuous with respect to $\mathbb{R}$), it follows that:

$$\begin{aligned}\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) &= \left|\sum_j \varphi\left(\frac{p_j}{r_j}\right) r_j - \sum_j \varphi\left(\frac{m_j}{r_j}\right) r_j + \sum_j \varphi\left(\frac{m_j}{r_j}\right) r_j - \sum_j \varphi\left(\frac{q_j}{r_j}\right) r_j\right| \\ &\leq \left|\sum_j \left(\varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{m_j}{r_j}\right)\right) r_j\right| + \left|\sum_j \left(\varphi\left(\frac{m_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right)\right) r_j\right| \\ &\leq \Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{M}) + \Delta_{\varphi,\mathbb{R}}(\mathbb{M},\mathbb{Q})\end{aligned}$$

As it is easy to verify, $\Delta_{\varphi,\mathbb{R}}(\cdot,\cdot)$ has the properties of a semi-distance, namely: symmetry, triangular inequality and semi-positivity ($\mathbb{P} = \mathbb{Q}$ then $\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = 0$ but $\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) = 0$ does not imply $\mathbb{P} = \mathbb{Q}$).

Remark 5 *By the previous definition, it is clear that the dissimilarity measure of any ,$\mathbb{P}$ with itself is given by $\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{P}) = 0$. Although this value seems intuitively natural, there can be circumstances for which this property is not necessary, even desired. In this case, it is possible to define the dissimilarity of an unspecified element with itself by:*

$$\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{R}) = \left|\sum_j \varphi\left(\frac{p_j}{r_j}\right) r_j - \sum_j \varphi\left(\frac{r_j}{r_j}\right) r_j\right| = \sum_j \varphi\left(\frac{p_j}{r_j}\right) r_j = \mathbb{E}^{\mathbb{R}}\left(\varphi\left(\frac{d\mathbb{P}}{d\mathbb{R}}\right)\right)$$

for every function φ such that : $\varphi(1) = 0$. This dissimilarity of $\mathbb{P}$ with itself is nothing else but the dissimilarity between $\mathbb{P}$ and $\mathbb{R}$.

Remark 6 *If $\mathbb{R}$ is a probability measure, one has:*

$$\begin{aligned}\Delta_{\varphi,\mathbb{R}}(\mathbb{P},\mathbb{Q}) &= \left|\sum_j \left(\varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right)\right) r_j\right| \\ &\leq \sum_j \left|\varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right)\right| r_j \\ &\leq \sup_j \left|\varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right)\right|\end{aligned}$$

Then the following quantities:

$$\sum_j \left| \varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right) \right| r_j$$

and:

$$\sup_j \left| \varphi\left(\frac{p_j}{r_j}\right) - \varphi\left(\frac{q_j}{r_j}\right) \right|$$

(or $\sup_j |\varphi(p_j) - \varphi(q_j)|$ *if* $\mathbb{R}$ *is a counting measure) are also as above, dissimilarity measures between the probability measures* $\mathbb{P}$ *and* $\mathbb{Q}$ *(with respect to* $\mathbb{R}$*). Depending on the chosen approach of a classification problem, one will use one of the three dissimilarity measures introduced just above.*

Remark 7 *The previous definitions and properties are of course, using usual modifications, valid if the set I is countable or continuous. One can see for instance Colin, Troupé et Vaillant [26]*

Being given two probability measures $\mathbb{P}_{e_1}$ and $\mathbb{P}_{e_2}$, where e_1 and e_2 are any two members of a finite set E, the dissimilarity measure between e_1 and e_2, noted by $\delta_{2,\varphi,\mathbb{R}}(e_1, e_2)$ or more simply by $\delta_2(e_1, e_2)$, is, by definition, given by:

$$\delta_2(e_1, e_2) = \Delta_{\varphi,\mathbb{R}}\left(\mathbb{P}_{e_1}, \mathbb{P}_{e_2}\right)$$

which will be also noted, if there is no confusion, as $\Delta_\varphi\left(\mathbb{P}_{e_1}, \mathbb{P}_{e_2}\right)$. The mapping from $E \times E$ to $\mathbb{R}^+$ which associates to each pair (e_i, e_j) the positive number $\Delta_\varphi\left(\mathbb{P}_{e_i}, \mathbb{P}_{e_j}\right)$, satisfy the conditions of a binary dissimilarity measure as introduced in Diatta [1].

4.2.2 k-way dissimilarity measure

From a more general point of view, let us consider a non-empty subset $X = \{e_i\}$ of E where $1 \leq i \leq Card(X) \leq k$ and let $\delta_k(X)$ be the quantity

defined by:

$$\delta_k(X) = \sum_{i=1}^{Card(X)} \sum_{j \geq i} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_{e_j} \right) = \frac{1}{2} \sum_{i,j=1}^{\sharp(X)} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_{e_j} \right)$$

It is again easy to verify that this last quantity satisfies the definitions and the conditions as stated in Diatta [1] and therefore can be consider as a k-way dissimilarity measure. For instance, for a ternary dissimilarity measure, one has for every (x, y, z) from E^3 (if $\varphi(1) = 0$):

$$\delta_3(x, x, x) = 0 \ \ ; \ \ \delta_3(x, y, z) \geq 0$$

and:

$$\delta_3(x, y, z) = \delta_3(x, z, y) = \delta_3(y, x, z) = \delta_3(y, z, x) = \delta_3(z, x, y) = \delta_3(z, y, x)$$

Let us note that the first condition $\delta_3(x, x, x) = 0$ for every member x of X, is not really important, because, except for an additive constant, one can always reduce the problem to this case. However, in practice this condition is frequently fulfilled, particularly for most functions φ introduced in the first section.

In the same way, if one notes by $E^*_{\leq k}$ the family of nonempty subsets of E having at most k elements, then for every members X, Y of $E^*_{\leq k}$, such that $X \subseteq Y$, one has, as it can be easily check:

$$\delta_k(X) \leq \delta_k(Y)$$

Finally, if the data are displayed in a n-way contingency tables, it may happen that one wants to perform a classification of the random categorical variables for which, an estimation of the joint probability measure can be obtained. In this case, the mutual information is a k-way dissimilarity measure which

takes into account the stochastic dependance between the variables. Indeed, let us consider for a given value k, a subset $\widetilde{X}_p = (X_1, X_2, ..., X_p)$ of random variables ($p \leq k$) and let:

$$\delta_k(\widetilde{X}_p) = \mathfrak{I}_\varphi(X_1, X_2, ..., X_p)$$

be the mutual information between the variables $X_1, X_2, ..., X_p$. The properties of the mutual information as stated in the present section, allow us to easily verify that $\delta_k(\widetilde{X}_p)$ is really a k-way dissimilarity measure ($k \geq 2$) on the set of the random variables. Indeed, one has:

i) $\mathfrak{I}_\varphi(X_1, X_2, ..., X_p) \geq 0$ (if $\varphi(1) \geq 0$), where the equality holds if and only if the random variables $X_1, X_2, ..., X_p$ are stochastically independent.

ii) $\mathfrak{I}_\varphi(X_1, X_2, ..., X_p) \geq \mathfrak{I}_\varphi(X_1, X_2, ..., X_{p-1})$, where the equality holds if and only if the random variable X_p is stochastically independent from $X_1, X_2, ..., X_{p-1}$.

As an immediate consequence of the last inequality, it is obvious that if $\widetilde{X}_p$ is a subset of random variables from $\widetilde{X}_q$ with $p < q \leq k$ one has:

$$\delta_k(\widetilde{X}_p) \leq \delta_k(\widetilde{X}_q)$$

Remark 8 *Like in Diatta [1], one calls δ_k-ball, or more simply ball, of center X and radius $r \geq 0$, the subset of E, noted $B^{\delta_k}(X, r)$, defined by:*

$$B^{\delta_k}(X, r) = \{y \in E : \delta_k(X \cup \{y\}) \leq r\}$$

It will be noticed that, in accordance with the definition of a k-way dissimilarity, one must have $Card(X) \leq k - 1$. Therefore, the definition of $B^{\delta_k}(X, r)$ is meaningful only if $X \in X^*_{\leq k-1}$. In the same way, one calls δ_k-diameter, or more simply diameter, of any nonempty subset Z of E, the greatest dissimi-

larity, noted $diam_{\delta_k}(Z)$*, between members of* Z*. In other words:*

$$diam_{\delta_k}(Z) = \max_{T \in Z^*_{\leq k}} (\delta_k(T))$$

which can be reduced, by the isotonic property of the k*-way dissimilarity measure, to:*

$$diam_{\delta_k}(Z) = \max_{T \in Z^*_{=k}} (\delta_k(T))$$

where $Z^*_{=k}$ *stands for the family of subsets of* Z *having exactly* k *members. Finally, if* $X \in E^*_{\leq k}$*, the* (δ_k, k)*- ball, or more simply the* k*-ball relative to* δ_k*, generated by* X*, is the subset of* E*, noted* $B^{\delta_k}_X$*, given by:*

$$B^{\delta_k}_X = \begin{cases} B^{\delta_k}(X, \delta_k(X)) \quad if \;\; Card(X) \leq k-1 \\ \\ \bigcap\limits_{x \in X} B^{\delta_k}(X \backslash \{x\}, \delta_k(X)) \qquad otherwise \end{cases}$$

For instance, if δ_2 *is a binary dissimilarity measure, one has for* $X = \{x, y\}$*:*

$$B^{\delta_2}_X = B^{\delta_2}(x, \delta_2(x, y)) \cap B^{\delta_2}(y, \delta_2(x, y))$$

If $Card(X) \leq k-1$*, the* k*-ball* $B^{\delta_k}_X$*, relative to* δ_k *and generated by* X*, is given by:*

$$B^{\delta_k}_X = \{y \in E : \delta_k(X \cup \{y\}) \leq \delta_k(X)\}$$

However X*,* $(X \cup \{y\}) \in E^*_{\leq k}$ *with* $X \subseteq (X \cup \{y\})$*. It follows that for all members* y *of* $E \backslash X$ *belonging to* $B^{\delta_k}_X$*, one has:* $\delta_k(X) \leq \delta_k(X \cup \{y\}) \leq \delta_k(X)$ *and then* $\delta_k(X \cup \{y\}) = \delta_k(X)$*. Therefore, if* $B^{\delta_k}_X$ *includes some members* y *of* E *other than those of* X*, the previous equality shows that these members* y *are in a sense similar to at least one member of* X*. This is the case, for instance, for some* k*-way dissimilarity measures when* y *is the same as some members* x *of* X*. For every* $X \in E^*_{\leq k}$*, let us consider for a given* $\mathbb{R}$*, the* k*-way*

dissimilarity measure:

$$\delta_k(X) = \sum_{i=1}^{Card(X)} \sum_{j \geq i} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_{e_j} \right)$$

If $\mathbb{P}_y$ *is the probability measure associated to the member* y *of* E*, one has for* $Card(X) \leq k-1$*:*

$$\delta_k(X \cup \{y\}) = \sum_{i=1}^{Card(X)} \sum_{j \geq i} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_{e_j} \right) + \sum_{i=1}^{Card(X)} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_y \right) + \Delta_\varphi \left(\mathbb{P}_y, \mathbb{P}_y \right)$$

which, taking into account that $\Delta_\varphi \left(\mathbb{P}_y, \mathbb{P}_y \right) = 0$ *and that* $\delta_k(X) = \delta_k(X \cup \{y\})$*, gives:*

$$\sum_{i=1}^{Card(X)} \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_y \right) = 0$$

and therefore:

$$\forall i = 1, 2, ..., Card(X) \quad \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_y \right) = 0$$

or:

$$\forall i = 1, 2, ..., Card(X) \quad \delta_2(e_i, y) = \Delta_\varphi \left(\mathbb{P}_{e_i}, \mathbb{P}_y \right) = 0$$

In other words, for all $i = 1, 2, ..., Card(X)$*, the probability measures* $\mathbb{P}_y$ *and* $\mathbb{P}_{e_i}$ *are similar and it can be noticed that more* k *is large, more this condition is difficult to satisfy. In order to avoid some meaningless situations when* $y \in E \backslash X$ *(indeed, by the means of the triangular inequality, one will have in this case, for all* e_i, e_j *in* X *and for all* y *in* $B_X^{\delta_k} \backslash X$*:* $\delta_2(e_i, e_j) \leq \delta_2(e_i, y) + \delta_2(y, e_j) = 0$*, and then* $\delta_2(e_i, e_j) = 0$*), one has to consider the case where* y *is identical to one of the members of* X *(*x *and* y *correspond to the same probability measure). It is then easy to verify that:*

$$diam_{\delta_k}(B_X^{\delta_k}) = \delta_k(X)$$

If $Card(X) = k$*, one has:*

$$B_X^{\delta_k} = \bigcap_{x \in X} B^{\delta_k}(X \backslash \{x\}, \delta_k(X))$$

Therefore, if $\mathbb{P}_x$ *is the probability measure associated to the member* x *of* X*, the equality* $\Delta_\varphi(\mathbb{P}_x, \mathbb{P}_x) = 0$*, leads to:*

$$\delta_k(X\backslash\{x\}) = \delta_k(X) - \sum_{i=1}^{k-1} \Delta_\varphi(\mathbb{P}_{e_i}, \mathbb{P}_x)$$

Thus, for every member y *of* $B^{\delta_k}(X\backslash\{x\}, \delta_k(X))$ *and for every member* x *of* X*, one will have:*

$$\delta_k((X\backslash\{x\}) \cup \{y\}) = \delta_k(X\backslash\{x\}) + \sum_{i=1}^{k-1} \Delta_\varphi(\mathbb{P}_{e_i}, \mathbb{P}_y) \leq \delta_k(X)$$

Consequently, y *being identical with a member* x *of* X *(to avoid again trivial cases) one has:*

$$diam_{\delta_k}(B_X^{\delta_k}) = \delta_k(X)$$

5 Conclusion

We proposed a way to specify the Galois k-weak hierarchy associated with a k-way dissimilarity function, in a probabilistic meet-closed description context. The members of such a Galois k-weak hierarchy are pairs of the form $(X, \wedge\delta(X))$, where X is a Galois closed weak cluster associated with the given k-way dissimilarity function. On the other hand, we placed ourselves in the framework of a meet-closed description context where entity descriptions are probability distributions. Then we showed that the expectation is a strict valuation when the entity description space is endowed with the stochastic order.

References

[1] J. Diatta, Description-meet compatible multiway dissimilarities, Discrete Applied Mathematics 154 (2006) 493–507.

[2] P. Bertrand, M. F. Janowitz, The k-weak hierarchical representations: an extension of the indexed closed weak hierarchies, Discrete Applied Mathematics 127 (2003) 199–220.

[3] H.-J. Bandelt, A. W. M. Dress, An order theoretic framework for overlapping clustering, Discrete Mathematics 136 (1994) 21–37.

[4] J. Diatta, Dissimilarités multivoies et généralisations d'hypergraphes sans triangles, Math. Inf. Sci. hum. 138 (1997) 57–73.

[5] A. Soule, K. Salamatian, N. Taft, R. Emilion, K. Papaginiaki, Flow classification by histograms, in Sigmetrics'04, http://www-rp.lip6.fr/ soule/SiteWeb/Publication.php (2004).

[6] H.-J. Bandelt, A. W. M. Dress, Weak hierarchies associated with similarity measures: an additive clustering technique, Bull. Math. Biology 51 (1989) 113–166.

[7] M. Barbut, B. Monjardet, Ordre et classification, Hachette, Paris, 1970.

[8] E. Diday, R. Emilion, Maximal and stochastic Galois lattices, Discrete Applied Mathematics .

[9] A. Muller, M. Scarsini, Stochastic order relations and lattices of probability measures, http://web.econ.unito.it/scarsini/psfiles/lattice20040617.pdf (2004).

[10] F. Baccelli, P. Brémaud, Elements of queing theory, Springer, 2003.

[11] H. Bandemer, W. Näther, Fuzzy data analysis, Kluwer academic publishers, 1992.

[12] I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, Studia Scientiarum Mathematicarum Hungarica 2 (1967) 299–318.

[13] I. Csiszár, A class of measures of informativity of observation channels, Periodica Mathematica Hungarica 2 (1972) 191–213.

[14] I. Csiszár, Information measures : A critical survey, in: Transaction of the seventh Prague Conference on Information theory Statistical Decision functions Random Processes, Vol. A, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1977, pp. 73–86.

[15] S. Ali, S. Silvey, A general class of coefficients of divergence of one distribution from another, J.Roy.Statist.Soc. B.28 (1966) 131–142.

[16] J. Zakai, M. Ziv, On functionnals satisfying a data-processing theorem, IEEE Transactions IT-19 (1973) 275–282.

[17] P. Goël, Information measures and bayesian hierarcichal models, Tech. Rep. 81-41, Departement of Statistics, Purdue University, West Lafayette (1981).

[18] B. Adhikari, D. Joshi, Distance, discrimination et résumé exhaustif, Publications de l'Institut de Statistique de l'Université de Paris 5 (1956) 57–74.

[19] J. Aczél, Z. Daróczy, On measures of information and their characterizations, Academic Press, New York, 1975.

[20] A. Rényi, On measures of entropy and information, in: Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, 1961, pp. 547–561.

[21] M. Pinsker, Information and information stability of random variables and processes, Holden-Day, 1964.

[22] R. McEliece, The theory of information coding, Encyclopedia of mathematics and its applications, Addison Wesley, 1977.

[23] M. Gavurin, On the value of information, Selected Translations in Mathematical Statistics and Probability 7 (1968) 193–202.

[24] R. Sibson, Information radius, Z. Wahrsch' theorie & verw. Geb. 14 (1969) 149–160.

[25] N. Jardine, R. Sibson, Mathematical Taxonomy, Wiley, New York, 1971.

[26] B. Colin, J. Vaillant, M. Troupé, Information mutuelle et divergence : estimation, codage et classification, Tech. Rep. 11, Département de Mathématiques, Université de Sherbrooke (Québec), http://www.usherbrooke.ca/mathematiques/telechargement/ (2004).