Entropies and Co–entropies for Incomplete Information Systems*

Daniela Bianucci, Gianpiero Cattaneo, and Davide Ciucci

Dipartimento di Informatica, Sistemistica e Comunicazione
Università di Milano – Bicocca
Via Bicocca degli Arcimboldi 8, I–20126 Milano, Italia
{bianucci,cattang,ciucci}@disco.unimib.it

Abstract. A partitioning approach to the problem of dealing with the entropy of incomplete information systems is explored. The aim is to keep into account the incompleteness and at the same time to obtain a probabilistic partition of the information system. For the resulting probabilistic partition, measures of entropy and co–entropy are defined, similarly to the entropies and co–entropies defined for the complete case.

Keywords: entropy, co–entropy, incomplete information system.

1 Introduction: Qualitative and Quantitative Valuations of Roughness for Complete Information Systems

In this work, we discuss the entropy of incomplete information systems as an extension of the approach based on partitions from complete information systems. In order to introduce an approach of *probability partition* from an incomplete information system, let us first recall how one gets a partition from a complete information system, and thus how one can apply a measure of rough entropy when dealing with an information system.

Let us recall that the original Pawlak approach to rough sets is essentially based on an *approximation space*, i.e., a pair $\langle X, \pi \rangle$ where X is a (finite) set, called the *universe* of *objects*, and $\pi = \{A_1, A_2, \dots, A_N\}$ is a partition of X, in general induced by the indistinguishability equivalence relation from a complete information system [1]. The subsets A_j are the *elementary sets* (or also *events*), each of which can be interpreted as a *granule of knowledge* supported by the partition. We denote by $gr_\pi(x)$ the granule (equivalence class) from π which contains the point $x \in X$. In the rough set theory, once fixed a partition π of X, any of its subsets H can be approximated from the bottom and from the top by the two *lower* and *upper approximations* defined respectively as: $l_\pi(H) := \cup\{A_i \in \pi : A_i \subseteq H\}$ and $u_\pi(H) := \cup\{A_j \in \pi : H \cap A_j \neq \emptyset\}$, producing the rough approximation of H defined as the pair $r_\pi(h) = (l_\pi(H), u_\pi(H))$ (with trivially $l_\pi(H) \subseteq H \subseteq u_\pi(H)$), see [2] for a complete discussion. We can also

* The author's work has been supported by MIUR\PRIN project "Automata and Formal languages: mathematical and application driven studies."

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 84–92, 2007.

define the *boundary region* of H as $b_\pi(H) = u(H) \setminus l(H)$, and its *external region* as $e_\pi(H) = X \setminus u(H)$. Obviously, whatever be the starting original partition π, for any subset H the triple $\pi(H) = \{l_\pi(H), b_\pi(H), \ e_\pi(H)\}$ is a new partition of X, which depends from the choice of the subset H.

These considerations can be applied to the case of a complete Information System (IS), formalized by a triple $IS := \langle X, Att, F \rangle$ consisting of a nonempty finite set X of objects, a nonempty finite set of attribute Att, and a mapping $F : X \times Att \to V$ which assigns to any object $x \in A$ the value $F(x, a)$ assumed by the attribute $a \in Att$ [1,3,4]. Indeed, in this IS case the partition generated by a set of attributes $\mathcal{A}$, denoted by $\pi_{\mathcal{A}}(IS)$, consists of equivalence classes of *indistinguishable* objects A_i, i.e., two objects $x, y \in A_i$ iff for any attribute $a \in \mathcal{A}$, the condition $F(x, a) = F(y, a)$ holds.

In many applications it is of a certain interest to analyze the variations occurring inside two information systems labelled with two parameters t_1 and t_2. In particular, one has to do mainly with the following two cases in both of which the set of objects remains invariant:

(1) *dynamics* (see [5]), in which $IS_{t_1} = (X, Att_1, F_1)$ and $IS_{t_2} = (X, Att_2, F_2)$ are under the conditions that $Att_1 \subset Att_2$ and $\forall x \in X$, $\forall a_1 \in Att_1$: $F_2(x, a_1) = F_1(x, a_1)$. This situation corresponds to a dynamical increase of knowledge (t_1 and t_2 are considered as time parameters, with $t_1 < t_2$) for instance in a medical database the increase corresponds to the fact that during the researches on the disease some symptoms which have been neglected at time t_1 become relevant at time t_2 under some new investigations.

(2) *reduct*, in which $IS_{t_1} = (X, Att_1, F_1)$ and $IS_{t_2} = (X, Att_2, F_2)$ are under the conditions that $Att_2 \subset Att_1$ and $\forall x \in X$, $\forall a_2 \in Att_2$: $F_2(x, a_2) = F_1(x, a_2)$. In this case it is of a certain interest to verify if the corresponding partitions are invariant $\pi_{Att_2}(IS_{t_2}) = \pi_{Att_1}(IS_{t_1})$, or not.

From the point of view of the rough approximations of subsets Y of the universe X, both these cases can be treated under a unified formal framework in which during the time evolution $t_1 \to t_2$ one try to relate the corresponding variation of partitions $\pi_{t_1} \to \pi_{t_2}$ with, for instance, the boundary transformation $b_{t_1}(Y) \to b_{t_2}(Y)$. First of all, as to the partitions of X, whose collection will be denoted by $\Pi(X)$, their more interesting structure is the one of complete lattice (see [6]) with respect to the partially order relation $\pi_1 \preceq \pi_2$, which can be formalized in one of the following mutually equivalent forms: (por1) $\forall A \in \pi_1$, $\exists B \in \pi_2$: $A \subseteq B$; (por2) $\forall B \in \pi_2$, $\exists \{A_{i_1}, A_{i_2}, \ldots, A_{i_h}\} \subseteq \pi_1$: $B = A_{i_1} \cup A_{i_2} \cup \ldots \cup A_{i_h}$; (por3) $\forall x \in X$, $gr_{\pi_1}(x) \subseteq gr_{\pi_2}(x)$ (as shown in [7], an extension of these three formulations to the case of coverings leads to different binary relations of quasi–orderings). The lattice $\Pi(X)$ of all partitions of X is lower bounded by the least element $\pi_d := \{\{x\} : x \in X\}$ (the *discrete* partition) consisting of all singletons from X, and the greatest element $\pi_t := \{X\}$ (the *trivial* partition) whose unique equivalence class is the whole universe. If $\pi_1 \preceq \pi_2$ we say that π_1 (resp., π_2) is *finer* (resp., *coarser*) than π_2 (resp., π_1). The induced *strict ordering* on partition, denoted by $\pi_1 \prec \pi_2$, is defined as $\pi_1 \preceq \pi_2$ and $\pi_1 \neq \pi_2$. This means that it must exists at least an equivalence class $B_i \in \pi_2$

such that its partition with respect to π_1 is formed at least of two subsets, i.e., $\exists\{A_{i_1}, A_{i_2}, \dots, A_{i_p}\} \subseteq \pi_1$, with $p \geq 2$, s.t. $B_i = A_{i_1} \cup A_{i_2} \cup \dots \cup A_{i_p}$.

Let us note that if $\pi_1 \preceq \pi_2$, then the two rough approximations of a given subset Y, $r_{\pi_i}(Y) = (l_{\pi_i}(Y), u_{\pi_i}(Y))$, for $i = 1, 2$, are such that $l_{\pi_2}(Y) \subseteq l_{\pi_1}(Y) \subseteq Y \subseteq u_{\pi_1}(Y) \subseteq u_{\pi_2}(Y)$, i.e., the rough approximation of Y with respect to the partition π_1 is *better* than the rough approximation of the same subset with respect to π_2. This leads to a first but only *qualitative* valuation of the *roughness* of a subset Y of the universe expressed by the law: $\pi_1 \preceq \pi_2$ implies that for $\forall Y$, $b_{\pi_1}(Y) \subseteq b_{\pi_2}(Y)$. The delicate point is that the condition of strict ordering $\pi_1 \prec \pi_2$ does not assure that for $\forall Y$, $b_{\pi_1}(Y) \subset b_{\pi_2}(Y)$. It is possible to give some very simple counter–examples in which notwithstanding $\pi_1 \prec \pi_2$ one has that $\exists Y_0$: $b_{\pi_1}(Y_0) = b_{\pi_2}(Y_0)$ [8,7], and this is not a desirable behavior of such a qualitative valuation of roughness. On the other hand, in many practical applications (for instance in the attribute reduction procedure), it is interesting not only to have a possible qualitative valuation of the roughness of a generic subset Y, but also a *quantitative* valuation formalized by a mapping $E : \Pi(X) \times 2^X \to [0, 1]$ assumed to satisfy (at least) the following two minimal requirements:

(re1) the *strict monotonicity condition*: for any $Y \in 2^X$, $\pi_1 \prec \pi_2$ implies $E(\pi_1, Y) < E(\pi_2, Y)$;

(re2) the *boundary conditions*: for $\forall Y \in 2^X$, $E(\pi_d, Y) = 0$ and $E(\pi_t, Y) = 1$.

In the sequel, sometimes we will use the notation $E_\pi : 2^X \to [0, 1]$ to denote the above mapping in which the partition $\pi \in \Pi(X)$ is considered fixed once for all. The interpretation of condition (re2) is possible under the assumption that a quantitative valuation of the roughness $E_\pi(Y)$ should be *directly* related to its boundary by $|b_\pi(Y)|$. From this point of view, the value 0 corresponds to the discrete partition for which the boundary of any subset Y is empty, and so its rough approximation is $r_{\pi_d}(Y) = (Y, Y)$ with $|b_{\pi_d}(Y)| = 0$, i.e., a crisp situation. On the other hand, the value 1 corresponds to the trivial partition in which the boundary of any nontrivial subset $Y (\neq \emptyset, X)$ is the whole universe, and so its rough approximation is $r_{\pi_t}(Y) = (\emptyset, X)$ with $|b_{\pi_t}(Y)| = |X|$, i.e., the minimum of sharpness or maximum of roughness.

This being stated, in literature one can find a lot of quantitative *measures of roughness* of Y relatively to a given partition $\pi \in \Pi(X)$ formalized as mappings $\rho_\pi : 2^X \to [0, 1]$. The *accuracy* of the set Y with respect to the partition is then defined as $\alpha_\pi(Y) = 1 - \rho_\pi(Y)$. Two of the more interesting roughness measures are $\rho_\pi^{(P)}(Y) := \frac{|b_\pi(Y)|}{|u_\pi(Y)|}$ [3] and $\rho_\pi^{(C)}(Y) := \frac{|b_\pi(Y)|}{|X|}$ [7]. These roughness measures satisfy the above "boundary" condition (re2), but their drawback is that the strict condition on partitions $\pi_1 \prec \pi_2$ does not assure a corresponding strict behavior $\forall Y$, $b_{\pi_1}(Y) \subset b_{\pi_2}(Y)$, and so also the strict correlation $\rho_{\pi_1}(Y) < \rho_{\pi_2}(Y)$ cannot be inferred. In other words, in general a rough measure is monotonic, but not strictly monotonic, contrary to the above requirement (re1).

This drawback can be overcome according to at least two strategies: either by some new strictly monotonic roughness measures or maintaining one of the monotonic roughness measures ρ_π and considering a strict monotonic function

$\Omega : \Pi(X) \to [0,1]$ in such a way that the new mapping $E(\pi, Y) := \rho_\pi(Y) \cdot \Omega(\pi)$ turns out to be strictly monotonic. In this paper we explore this second possibility in which, for the sake of simplicity, the required function is not the normalized Ω but it is given by a co–entropy function (also granularity measure) $E : \Pi(X) \to [0, k]$, where k is a suitable constant, and from which it is possible to induce the normalized $\Omega(Y) = E(Y)/k$. This is discussed in the following section.

2 Global and Pointwise Entropies and Co–entropies from Partitions

Given a partition $\pi = (A_1, A_2, \ldots, A_N)$ of the universe X by the *elementary events* A_i, one can construct the σ–algebra $\mathcal{E}(\pi)$ of *events* generated by π consisting of the empty set and all the set theoretic unions of elementary events. In the measurable space $(X, \mathcal{E}(\pi))$ the *counting measure* $m_\pi : \mathcal{E}(\pi) \to \mathbb{R}_+$ assigns to any event E the corresponding measure $m_\pi(E) = |E|$ (its cardinality). In this space with measure $(X, \mathcal{E}(\pi), m_\pi)$ we can introduce the vector $\boldsymbol{m}(\pi) = (m_\pi(A_1), m_\pi(A_2), \ldots, m_\pi(A_N)$, with $m_\pi(A_j) > 0$ for every j and $\sum_{j=1}^{N} m_\pi(A_j) = |X|$, which is a measure of the granulation, called the *granularity distribution* induced by π. Finally, it is possible to introduce the vector $\boldsymbol{p}(\pi) := (p_\pi(A_1), p_\pi(A_2), \ldots, p_\pi(A_N))$, where each $p_\pi(A_j) := \frac{|A_j|}{|X|}$ represents the probability of occurrence of the granule A_j. Since for each j we have that $p_\pi(A_j) > 0$ and $\sum_{j=1}^{N} p_\pi(A_j) = 1$, the vector $\boldsymbol{p}(\pi)$ constitutes a *probability distribution* induced by granulation. This being stated, in this section we consider the following two quantities depending from the partition π:

$$E(\pi) = \frac{1}{|X|} \sum_{i=1}^{N} |A_i| \log |A_i| \tag{2.1a}$$

$$H(\pi) := -\sum_{j=1}^{N} p(A_j) \cdot \log p(A_j) = -\frac{1}{|X|} \sum_{i=1}^{N} |A_i| \log \frac{|A_i|}{|X|}. \tag{2.1b}$$

Let us note that $E(\pi)$ depends only from the *granularity distribution* $\boldsymbol{m}(\pi)$, whereas $H(\pi)$ depends from the *probability distribution* $\boldsymbol{p}(\pi)$. In our opinion this leads to two different semantical interpretations of these quantities. Indeed, in agreement with the *information theory*, since the granule A_j has probability $p(A_j)$, we shall say that the quantity $I[p(A_j)] := -\log p(A_j)$ is the *uncertainty* associated with the granule A_j. Thus, the quantity $H(\pi)$, as expectation of the discrete random variable $I[p(A_j)]$ with probability $p(A_j)$, is the *average uncertainty* relatively to the probability distribution $\boldsymbol{p}(\pi)$, i.e., it measures the *uncertainty* of the granulation. According to Shannon [9], $H(\pi)$ is called the *entropy* of the partition π. Besides this entropy, the quantity $E(\pi)$ can be defined as *co–entropy* owing to its complementarity role with respect to the entropy $H(\pi)$ formalized by the identity $E(\pi) + H(\pi) = \log |X|$, whatever be the partition π. Let us note that in [10] this quantity has been called *measure of the granularity* since it "is basically an expectation of granularity with respect to all subsets in a partition". The following strict monotonic (resp., anti–monotonic) behavior of co–entropy (resp., entropy) is a standard result (for the entropy see for instance [11]): $\pi_1 \prec \pi_2$ implies $E(\pi_1) < E(\pi_2)$ and $H(\pi_2) < H(\pi_1)$. Since the trivial (resp., discrete) partition π_t (resp., π_d) is the greatest (resp., least) element of the lattice of all partitions, it is easy to see that for any partition π it is $0 = E(\pi_d) \le E(\pi) \le E(\pi_t) = \log |X|$, according to the fact that "the coarsest partition π_t has the maximum granularity value $\log |X|$ and the finest partition π_d has the minimum granularity value 0" [10]. So the required normalized co–entropy (granular entropy) is $\Omega(\pi) := \frac{E(\pi)}{\log |X|}$.

Note that in [12] the entropy $H(\pi)$, with the corresponding anti–monotonic behavior, has been assumed as a "measure of granularity", but this (formally legitimate) choice is in contrast with the strict monotonicity (meta–)requirement subsumed by $E(\pi, Y)$ as "local" (i.e., depending from Y) measure of roughness. Further, this choice suffers also of another drawback. As previously underlined, in information theory the entropy $H(\pi)$ is interpreted as a measure of the uncertainty of the probability distribution generated by the partition π. In conclusion, the different behaviors of $H(\pi)$ and $E(\pi)$ with respect to the variation of the partition π lead to different semantics: $H(\pi)$ can be interpreted as a measure of the *information uncertainty*, $E(\pi)$ as a measure of *partition granularity*, and $\rho_\pi(Y) \cdot E(\pi)$ as a *local* measure of *rough granularity*. The finer is the partition and the greatest (resp., lower) is the uncertainty (resp., the roughness).

In order to appreciate a possible generalization of these arguments to the case of incomplete IS, for instance according to the approach of [13], in [7] it has been introduced also in the partition context the new notions of *pointwise* entropy and co–entropy as the two mappings in which the sum involves the "local" information given by all the equivalence classes $gr(x)$, with corresponding "probabilities" $\mu_\pi(x) = \frac{|gr(x)|}{|X|}$, for the object x ranging on the universe X:

$$E_{LX}(\pi) = \frac{1}{|X|} \sum_{x \in X} |gr(x)| \cdot \log |gr(x)| = \frac{1}{|X|} \sum_{i=1}^{N} |A_i|^2 \cdot \log |A_i| \tag{2.2a}$$

$$H_{LX}(\pi) = -\sum_{x \in X} \mu_\pi(x) \cdot \log \mu_\pi(x) = -\frac{1}{|X|} \sum_{i=1}^{N} |A_i|^2 \cdot \log \frac{|A_i|}{|X|} \tag{2.2b}$$

Trivially, $\forall \pi \in \Pi(X)$, $0 \le E(\pi) \le E_{LX}(\pi)$. In the sequel, we refer to $E(\pi)$ as the *global* entropy and to $E_{LX}(\pi)$ as the *pointwise* one. Moreover, setting $\mu(\pi) := \sum_{x \in X} \mu(x)$, one gets that $E_{LX}(\pi) + H_{LX}(\pi) = \log |X| \cdot \mu(\pi)$, with this latter depending on the partition π. Note that the probability vector $\boldsymbol{p}_{LX} := (\mu_\pi(x_1), \mu_\pi(x_2), \ldots, \mu_\pi(x_{|X|}))$ is not a probability distribution since the sum of its components is $\mu(\pi) \ge 1$. Notwithstanding this drawback, from (2.2a) it follows that the strict monotonicity condition holds also for the pointwise co–entropy: $\pi_1 \prec \pi_2$ implies $E_{LX}(\pi_1) < E_{LX}(\pi_2)$. Of course, in this case one has that for any π: $0 \le E_{LX}(\pi) \le |X| \cdot \log |X|$, with corresponding normalized co–entropy (granulation measure) $\Omega_{LX}(\pi) := \frac{E_{LX}(\pi)}{|X| \cdot \log |X|}$. Unfortunately, H_{LX} presents *neither* monotonic *nor* anti–monotonic behavior.

3 Incomplete Information Systems and Definition Domain

An *incomplete* information system is formalized as a triple $\langle X, Att, F\rangle$ where F is a mapping *partially* defined on a subset $\mathcal{D}(F)$ of $X \times Att$ under the following two *non–redundancy* conditions: (1) about objects: for every object $x \in X$ there exists at least an attribute $a \in Att$ such that $(x, a) \in \mathcal{D}(F)$; (2) about attributes: for every attribute $a \in Att$ there exists at least an object $x \in X$ such that $(x, a) \in \mathcal{D}(F)$. In this way also the mapping representation of an attribute a is partially defined on the *definition domain* $X_a := \{x \in X : (x, a) \in \mathcal{D}(F)\}$ of X (which is nonempty owing to the non–redundancy condition (2) about attributes) as the surjective mapping $f_a : X_a \mapsto val(a)$, where $val(a) := \{F(x, a) : x \in X_a\}$ is the set of all *possible values* of the attribute a. The non–redundancy condition (1) about objects assures that $\bigcup_{a\in Att} X_a = X$ (covering condition about attribute definition domains). Adding to $val(a)$ the further *null value* $*$, we obtain the new set $val^*(a)$ and it is possible to extend the partially defined mapping f_a to a global defined one, denoted by $f^*_a : X \mapsto val^*(a)$, which assigns to any object $x \in X$ the value $f^*_a(x) = f_a(x)$ if $x \in X_a$, and the value $f^*_a(x) = *$ otherwise.

Also in the case of incomplete information systems, if one fixes an attribute a and denotes by $\alpha_i \in val(a)$, the subset of the universe $A_i = f_a^{-1}(\alpha_i) = \{x \in X_a : f_a(x) = \alpha_i\}$ is the *elementary event* of all objects for which the attribute a assumes the value α_i. Further, for any family of attributes $\mathcal{A}$ one can construct the "common" definition domain $X_{\mathcal{A}} = \bigcup_{a\in\mathcal{A}} X_a$ and then it is possible to consider the multi–attributes mapping $f_{\mathcal{A}}$ assigning to any object $x \in X_{\mathcal{A}}$ the corresponding collection of values $f_{\mathcal{A}}(x) = (f^*_a(x))_{a\in\mathcal{A}}$, obtaining a mapping $f_{\mathcal{A}} : X_{\mathcal{A}} \mapsto val^*(\mathcal{A})$, with $val^*(\mathcal{A}) \subseteq \Pi_{a\in\mathcal{A}}\, val^*(a)$ the range of the mapping $f_{\mathcal{A}}$. Note that owing to the non–redundancy conditions for any $a \in Att$ at least one of the $f^*_a(x) \neq *$, and so $val^*(\mathcal{A})$ excludes the string consisting of all $*$. In order to extend to an incomplete information system the properties and considerations about entropy and co–entropy of partitions described at the end of section (2), we have at least two different possibilities [7].

(i) For any possible "value" $\boldsymbol{\alpha} \in val^*(\mathcal{A})$, one can construct the *granule* $f_{\mathcal{A}}^{-1}(\boldsymbol{\alpha}) = \{x \in X_{\mathcal{A}} : f_{\mathcal{A}}(x) = \boldsymbol{\alpha}\}$ of X labelled by $\boldsymbol{\alpha}$, also denoted by $[\mathcal{A}, \boldsymbol{\alpha}]$. The family of granules $gr(\mathcal{A}) = \{[\mathcal{A}, \boldsymbol{\alpha}] : \boldsymbol{\alpha} \in val^*(\mathcal{A})\}$ plus the *null granule* $[\mathcal{A}, *] = X \setminus X_{\mathcal{A}}$ (i.e., the collection of the objects in which all the attributes are unknown) constitutes a partition of the universe X, in which $gr(\mathcal{A})$ is a partition of the subset $X_{\mathcal{A}}$ of X (which can be considered as a "partial" partition of X).

(ii) Otherwise, we can consider the *covering* generated by a *similarity* (reflexive and symmetric, but in general nontransitive) relation. In the case of incompleteness it is often used the following relation [14]: two objects $x, y \in X$ are said to be *similar* if and only if $\forall a_i \in \mathcal{A} \subseteq Att$, either $f_{a_i}(x) = f_{a_i}(y)$ or $f_{a_i}(x) = *$ or $f_{a_i}(y) = *$.

The corresponding options are the following two. The first one, related to the above point (i) and investigated in this paper, involves partial partitions (related to "probabilistic partitions", i.e., partitions with respect to a measure m on events from X for which $m(X \setminus X_{\mathcal{A}}) = 0$). The second one, related to the point (ii), widely treated in literature by almost all the authors devoted to this argument (see for instance [13,15]), is applied to coverings [7]. The main point of difference, which gives to the approach (i) a real content of novelty, is that it is based on a generalization to probability partitions of the more economical *global* co–entropy (2.1a), whereas the approach (ii) generalizes the more complex *pointwise* co–entropy (2.2a) applied to coverings. Let us recall that in [7] different attempts has been investigated in order to give a global notion of co–entropy in the context of coverings, but all these attempts has been failed from the point of view of the monotonicity requirement. Finally, it is important that the results about incomplete ISs are not confused with the intrinsic arguments about complete ISs. These latter (as treated for instance in [16,12,10]), has to do with a narrow situation whose extension to the incomplete case is not trivial, and certainly original. This is what we discuss in the remaining part of the paper.

4 Entropies for Incomplete Information Systems

For any subset $\mathcal{A}$ of attributes of an incomplete information system, for the sake of simplicity, let us denote $val^*(\mathcal{A})$ as $V^*_{\mathcal{A}}$, for any $\boldsymbol{\alpha} \in V^*_{\mathcal{A}}$ the corresponding granule $f_A^{-1}(\boldsymbol{\alpha})$ as $A_{\boldsymbol{\alpha}}$ and let us set $X^*_{\mathcal{A}} = X \setminus X_{\mathcal{A}}$. Let us remark that the following holds: $x \notin X_{\mathcal{A}}$ iff $\forall a \in \mathcal{A} : f^*(x) = *$. Hence, the complementing (of $X_{\mathcal{A}}$ with respect to X) domain $X^*_{\mathcal{A}}$ is the collection of all states in which each attribute f_a of the family $\mathcal{A}$ is not defined (or in the information table the row corresponding to the object $x \in X^*_{\mathcal{A}}$ assumes the value $*$ in correspondence of any attribute $a \in \mathcal{A}$). From now on, if no confusion is likely, we simply use X^* instead of $X^*_{\mathcal{A}}$.

Now, we can define the measures $m_{\mathcal{A}}(A_{\boldsymbol{\alpha}}) = |A_{\boldsymbol{\alpha}}|$ and $m_{\mathcal{A}}(X^*) = 0$, and so $m_{\mathcal{A}}(X) = m_{\mathcal{A}}(\bigcup_{\boldsymbol{\alpha}} A_{\boldsymbol{\alpha}} \cup X^*) = \sum_{\boldsymbol{\alpha}} m_{\mathcal{A}}(A_{\boldsymbol{\alpha}}) + m_{\mathcal{A}}(X^*) = |X_{\mathcal{A}}|$, with the natural extension to the σ–algebra of events $\mathcal{E}_{\mathcal{A}}(X)$ from X generated by the elementary events $\{A_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in V^*_{\mathcal{A}}\} \cup \{A^* \in 2^X : A^* \subseteq X^*\}$ (with $m(A^*) = 0$), obtaining in this way a *finite measure* $m_{\mathcal{A}} : \mathcal{E}_{\mathcal{A}}(X) \to \mathbb{R}_+$ depending from the set of attributes $\mathcal{A}$. In particular, the measure of the whole universe changes with the choice of $\mathcal{A}$. The corresponding probabilities are then $p(A_{\boldsymbol{\alpha}}) = \frac{m_{\mathcal{A}}(A_{\boldsymbol{\alpha}})}{m_{\mathcal{A}}(X)} = \frac{|A_{\boldsymbol{\alpha}}|}{|X_{\mathcal{A}}|}$ and $p(X^*) = 0$. According to a widely used terminology, the collection $\pi(\mathcal{A}) = \{A_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in V^*_{\mathcal{A}}\}$ is a *probability partition* in the sense that the following hold: (1) each $p(A_{\boldsymbol{\alpha}}) > 0$; (2) $p(\bigcup_{\boldsymbol{\alpha}} A_{\boldsymbol{\alpha}}) = 1$; (3) $p(A_{\boldsymbol{\alpha}} \cap A_{\boldsymbol{\beta}}) = 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$.

Also in this case it is possible to define the *co–entropy* and the *entropy* of the probability partition generated by $\mathcal{A}$, similarly to (2.1a) and (2.1b), as follows:

$$E(\mathcal{A}) = \frac{1}{m_{\mathcal{A}}(X)} \sum_{\boldsymbol{\alpha} \in V^*_{\mathcal{A}}} m(A_{\boldsymbol{\alpha}}) \log m(A_{\boldsymbol{\alpha}}) \tag{4.1a}$$

$$H(\mathcal{A}) = -\sum_{\boldsymbol{\alpha} \in V^*_{\mathcal{A}}} p(A_{\boldsymbol{\alpha}}) \log p(A_{\boldsymbol{\alpha}}) = -\sum_{\boldsymbol{\alpha} \in V^*_{\mathcal{A}}} \frac{m_{\mathcal{A}}(A_{\boldsymbol{\alpha}})}{m_{\mathcal{A}}(X)} \log \frac{m_{\mathcal{A}}(A_{\boldsymbol{\alpha}})}{m_{\mathcal{A}}(X)} \tag{4.1b}$$

Trivially, $H(\mathcal{A})+E(\mathcal{A}) = \log|X_{\mathcal{A}}| = \log(m_{\mathcal{A}}(X))$, i.e., the non–negative quantity $E(\mathcal{A})$ "complements" the entropy $H(\mathcal{A})$ with respect to the value $\log(m_{\mathcal{A}}(X))$, which depends on the attribute collection $\mathcal{A}$. Let us remark that under the order condition $\mathcal{A} \subseteq \mathcal{B}$ on attributes we cannot state in general either $H(\mathcal{A}) \leq H(\mathcal{B})$ or $H(\mathcal{A}) \geq H(\mathcal{B})$. If for any collection $\mathcal{A}$ of attributes one defines the (globally normalized) probability $p^*(A_{\boldsymbol{\alpha}}) = \frac{|A_{\boldsymbol{\alpha}}|}{|X|}$, then the following definition can be given.

Definition 4.1. *Let $\langle X, Att, F\rangle$ be an incomplete information system, $\mathcal{A} \subseteq Att$ a collection of attributes, $X_{\mathcal{A}} \subseteq X$ the corresponding definition domain, and $\pi(\mathcal{A})$ the related pseudo–probability partition (pseudo since the condition (2) of probability partitions must be substituted by $p^*(\cup_{\boldsymbol{\alpha}} A_{\boldsymbol{\alpha}}) = |X_{\mathcal{A}}|/|X| \leq 1$).*

Then, we define the following co–entropy and entropy:

$$\tilde{E}(\mathcal{A}) := \left(\frac{|X| - |X_{\mathcal{A}}|}{|X|}\right) \log|X| + \frac{1}{|X|} \sum_{\boldsymbol{\alpha}\in V^*_{\mathcal{A}}} |A_{\boldsymbol{\alpha}}| \, \log|A_{\boldsymbol{\alpha}}| \tag{4.2a}$$

$$\tilde{H}(\mathcal{A}) := -\sum_{\boldsymbol{\alpha}\in V^*_{\mathcal{A}}} p^*(A_{\boldsymbol{\alpha}}) \log p^*(A_{\boldsymbol{\alpha}}) = -\frac{1}{|X|} \sum_{\boldsymbol{\alpha}\in V^*_{\mathcal{A}}} |A_{\boldsymbol{\alpha}}| \log\frac{|A_{\boldsymbol{\alpha}}|}{|X|} \tag{4.2b}$$

Also in this case we have that $\tilde{H}(\mathcal{A}) + \tilde{E}(\mathcal{A}) = \log|X|$. The following important result about monotonicity holds.

Theorem 4.1 (Monotonicity of $\tilde{H}(\mathcal{A})$). *Given an incomplete information system, let $\mathcal{A} \subseteq \mathcal{B}$ be two collections of attributes, and $\pi(\mathcal{B})$ and $\pi(\mathcal{A})$ the corresponding probability partitions. Then we have $\tilde{H}(\mathcal{A}) \leq \tilde{H}(\mathcal{B})$.*

Moreover, under the condition $|X_{\mathcal{B}}| > |X_{\mathcal{A}}|$ the following strict monotonicity holds: $\mathcal{A} \subset \mathcal{B}$ implies $\tilde{H}(\mathcal{A}) < \tilde{H}(\mathcal{B})$.

As a direct consequence of theorem 4.1, and making use of $\tilde{H}(\mathcal{A}) + \tilde{E}(\mathcal{A}) = \log|X|$, we have the following corollary regarding the co–entropy $\tilde{E}(\mathcal{A})$.

Corollary 4.1 (Anti–monotonicity of $\tilde{E}(\mathcal{A})$). *Let $\mathcal{A}$, $\mathcal{B}$ be two collections of attributes such that $\mathcal{A} \subseteq \mathcal{B}$. Then we have $\tilde{E}(\mathcal{B}) \leq \tilde{E}(\mathcal{A})$.*

5 Conclusions and Open Problems

We have illustrated a partitioning approach for incomplete information systems which take into account the incomplete nature producing at the same time a *probability* partition from one side (probability $p(A_{\boldsymbol{\alpha}}) = |A_{\boldsymbol{\alpha}}|/|X_{\mathcal{A}}|$) and a pseudo–probability partition on the other side (probability $p^*(A_{\boldsymbol{\alpha}}) = |A_{\boldsymbol{\alpha}}|/|X|$). We have then presented a definition of entropy and co–entropy for incomplete information systems based on the described partitioning approach.

We have shown that the entropy behaves monotonically and the co–entropy anti–monotonically, with respect to the collections of attributes. Let us stress that both the here defined co–entropies (4.1a) and (4.2a) result to be a generalization of the co–entropy (2.1a) of complete information systems.

The further step in this research will be the application of our co–entropy to the construction of reducts and rules in "real" information tables and the comparison, also from a computational point of view, with the "pointwise" co–entropy based on coverings considered in [7]. Indeed, even if the procedures to compute the here introduced co-entropy and the "pointwise" one are in the same complexity class, it can be easily seen that the former one always requires less operations than the last one.

References

1. Pawlak, Z.: Information systems - theoretical foundations. Information Systems **6** (1981) 205–218
2. Cattaneo, G.: Abstract approximation spaces for rough theories. [17] 59–98
3. Pawlak, Z.: Rough sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
4. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In Pal, S., Skowron, A., eds.: Rough Fuzzy Hybridization. Springer–Verlag, Singapore (1999) 3–98
5. Cattaneo, G., Ciucci, D.: Investigation about Time Monotonicity of Similarity and Preclusive Rough Approximations in Incomplete Information Systems. Volume 3066 of Lecture Notes in Artificial Intelligence., Springer–Verlag (2004) 38–48
6. Preparata, F.P., Yeh, R.T.: Introduction to Discrete Structures. Addison–Wesley Publ. Comp., Reading, Massachusetts (1973)
7. Bianucci, D., Cattaneo, G., Ciucci, D.: Entropies and co–entropies of coverings with application to incomplete information systems. Fundamenta Informaticae **75** (2007) 77–105
8. Beaubouef, T., Petry, F.E., Arora, G.: Information–theoretic measures of uncertainty for rough sets and rough relational databases. Journal of Information Sciences **109** (1998) 185–195
9. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423, 623–656
10. Yao, Y.: Probabilistic approaches to rough sets. Expert Systems **20** (2003) 287–297
11. Reza, F.M.: An Introduction to Information theory. Dover Publications, New York (1994) (originally published by Mc Graw-Hill, New York, 1961).
12. Duntsch, I., Gediga, G.: Roughian: Rough information analysis. International Journal of Intelligent Systems **16** (2001) 121–147
13. Liang, J., Xu, Z.: Uncertainty measure of randomness of knowledge and rough sets in incomplete information systems. Intelligent Control and Automata **4** (2000) 2526–2529 Proc. of the 3rd World Congress on Intelligent Control and Automata.
14. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Information Sciences **112** (1998) 39–49
15. Huang, B., He, X., Zhong, X.: Rough entropy based on generalized rough sets covering reduction. Journal of Software **15** (2004) 215–220
16. Gediga, G., Duntsch, I.: Rough approximation quality revisited. Artificial Intelligence **132** (2001) 219–234
17. Polkowski, L., Skowron, A., eds.: Rough Sets in Knowledge Discovery 1. Physica–Verlag, Heidelberg, New York (1998)