Abstract: The interestingness measures for pattern associations proposed in the data mining literature depend only on the observation of relative frequencies obtained from 2×2 contingency tables. They can be called "absolute measures". The underlying scale of such a measure makes statistical decisions difficult. In this paper we present the foundations and the construction of a probabilistic interestingness measure that we call likelihood of the link index. This enables to capture surprising association rules. Indeed, its underlying principle can be related to that of information theory philosophy; but at a relational level. Two facets are developed for this index: symmetrical and asymmetrical. Two stages are needed to build this index. The first is "local" and associated with the two single boolean attributes to be compared. The second corresponds to a discriminant extension of the obtained probabilistic index for measuring an association rule in the context of a relevant set of association rules. Our construction is situated in the framework of the proposed indices in the data mining literature. Thus, new measures have been derived. Finally, we designed experiments to estimate the relevance of our statistical approach, this being theoretically validated, previously.

Keywords: Probabilistic Intestingness Measure, Association Rules, Independence Random Model, Contingency tables.

A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link

Israël-César Lerman[1] and Jérôme Azé[2]

[1] Irisa-Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cédex
lerman@irisa.fr

[2] Laboratoire de Recherche en Informatique
Université Paris-Sud
91405 Orsay Cédex
aze@lri.fr

1 Introduction

Seeking for a relevant interestingness measure in the context of a given data base is a fundamental task at the heart of data mining problems. We assume that the data are given by a a set of objects described by a set of boolean attributes. Let us denote them by $\mathcal{O}$ and $\mathcal{A}$, respectively. The crossing between $\mathcal{O}$ and $\mathcal{A}$ leads to an incidence table. This indicates for each object the subset of attributes (properties) associated with its definition. Let us denote by n and p the cardinalities of $\mathcal{O}$ and $\mathcal{A}$.

Let α_i^j specify the value of the attribute a^j for the object o_i: $\alpha_i^j = a^j(o_i), 1 \leq i \leq n, 1 \leq j \leq p$. The two possible values for a given α_i^j are "true" and "false". Generally the "true" and the "false" values are coded by the numbers 1 and 0, respectively. Without loss of generality we may suppose a "true" value as more significant than a "false" one. This is generally expressed in terms of statistical frequencies: the number of objects where a given attribute is "true" is lower than that where it is "false". We will represent a boolean attribute a by its extension $\mathcal{O}(a)$ which represents the subset of objects where a is true. Thus, $\mathcal{A}$ is represented by a set of parts of the object set $\mathcal{O}$.

The set $\mathcal{O}$ is generally obtained from a training set provided from a universe of objects $\mathcal{U}$. On the other hand, the boolean attribute set $\mathcal{A}$ can be obtained from conjunctions of more elementary attributes, called itemsets.

Determining "significant" itemsets is a crucial problem of "Data Mining" [1, 6]. However, this problem will not be addressed here.

Let us now introduce some notations. With the attribute a of $\mathcal{A}$ we associate the negated attribute $\neg a$ that we represent by the complementary subset of $\mathcal{O}(a)$ in $\mathcal{O}$. For a given pair (a, b) of $\mathcal{A} \times \mathcal{A}$, we introduce the conjunctions $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$ and $\neg a \wedge \neg b$ that are respectively represented by $\mathcal{O}(a) \cap \mathcal{O}(b)$, $\mathcal{O}(a) \cap \mathcal{O}(\neg b)$, $\mathcal{O}(\neg a) \cap \mathcal{O}(b)$ and $\mathcal{O}(\neg a) \cap \mathcal{O}(\neg b)$. Cardinalities of these sets are respectively denoted by $n(a \wedge b)$, $n(a \wedge \neg b)$, $n(\neg a \wedge b)$ and $n(\neg a \wedge \neg b)$. Finally, $n(a)$ and $n(b)$ will designate the set cardinalities of $\mathcal{O}(a)$ and $\mathcal{O}(b)$. These cardinalities appear in the contingency table crossing the two binary attributes $\{a, \neg a\}$ and $\{b, \neg b\}$. Consider also the ratios of these cardinalities over the number of objects n. These define the following relative frequencies or proportions: $p(a \wedge b)$, $p(a \wedge \neg b)$, $p(\neg a \wedge b)$ and $p(\neg a \wedge \neg b)$.

Relative to the entire set of attribute pairs $\mathcal{A} \times \mathcal{A}$ the objective consists in setting up a reduced subset of pairs (a^j, a^k), $1 \leq j < k \leq p$, such that a true value for the attribute a^j has a real tendency to imply a true value for that a^k. In order to measure numerically such a tendency that we denote by $(a \rightarrow b)$ for the ordered pair of boolean attributes (a, b), many association coefficients have been proposed. Inspired by different metric principles, they have not necessarily comparable behaviours for pattern association in a given application domain. Methodological comparisons between these measures are provided in the most recent research works [14, 16, 27]. Logical, statistical and semantical facets of a collection of 15 interestingness measures are analyzed in [14]. Comparison behaviour study of 20 indices is considered in [16]. Pairwise indices are compared according to the similarity of the rankings that they determine on a set of rules. Moreover, in this contribution eight formal criteria are considered to characterize in a global manner the properties of a measure. The desired properties proposed in [27] are substantially different from the latter ones. Relative to an ordered pair of boolean attributes (a, b) belonging to $\mathcal{A} \times \mathcal{A}$, these properties are more local and directly associated with a transformation of the respective entries of the contingency table crossing $\{a, \neg a\}$ with $\{b, \neg b\}$ that we have introduced above. In the invariance properties considered in [27], the studied interestingness measures are taken one by one. However, some investigation about the transformation of one type of measure to another, is required. This aspect is considered in section 4, where we focus more particularily on the Confidence, Loevinger and Gras's entropic indices.

Most of the interestingness measures take into account each attribute pair independently. The formulation of such a given measure M for an ordered pair (a, b) depends only on the above mentioned proportions. M is considered as an *absolute measure*. For a couple $\{(a, b), (a', b')\}$ of attribute pairs, the underlying numerical scale enables to answer the question of determining the most stressed association between $(a \to b)$ and $(a' \to b')$. However, without loss of generality, by assuming

$$M(a \to b) > M(a' \to b') \tag{1}$$

we cannot evaluate how much the intensity of $(a \to b)$ is significantly greater than that of $(a' \to b')$. Moreover, for a given ordered pair (a, b) occurring in two different data bases, it is difficult to situate comparitively by means of an absolute measure M the intensity evaluation in each of both data bases. In particular, the size value n does not intervene in the mathematical expression of M.

Our method consists of evaluating in a relative way the degree of implication $a \to b$ by using an original notion of probabilistic index measuring how much unlikely in terms of probability the pattern is strengthened. A random model of no relation or independence is introduced associating with the observed incidence table a random one. Let us denote by $\mathcal{N}$ this random model. Then the general idea consists of substituting the initial scale given by M for that corresponding to the following equation:

$$P^{\mathcal{N}}(a \to b) = Prob\{M(a^* \to b^*) \leq M(a \to b) \mid \mathcal{N}\} \tag{2}$$

where (a^*, b^*) is an ordered pair of independent random attributes associated with (a, b) according to the model $\mathcal{N}$. There are two forms of this random model. The first one that we call "context free" model is local and does only concern the observed ordered pair (a, b) of boolean attributes (see section 2 and subsection 4.2). The second form qualified by "in the context" takes into account mutual comparison between all the attribute pairs or a relevant part of them (see section 3 and subsection 4.3). The second version of the random model $\mathcal{N}$ has a conditional meaning. For this model and more precisely, the above measure M is replaced by a standardized version M_s with respect to a relevant subset of $\mathcal{A} \times \mathcal{A}$ (see section 3 and subsection 4.3). In these conditions, the likelihood of the link probabilistic index can be written:

$$P_s^{\mathcal{N}}(a \to b) = Prob\{M_s(a^* \to b^*) \leq M_s(a \to b) \mid \mathcal{N}\} \tag{3}$$

Such a probabilistic interestingness measure provides clearer answers to the above evaluation questions. Moreover, in the context of a given data base a threshold value filtering the strongest rules is more easily controlled.

Thus our approach refers to the philosophy of the information theory, but at the level of the observed mutual relations. We need to quantify interesting implicative events by means of an index associated with a probability scale. The valuation of a given event, defining an association rule, takes into account all its possible positions relatively to an interesting potential subset of association rules associated in the context of $\mathcal{A} \times \mathcal{A}$.

In fact, for a given form of the independence relation $\mathcal{N}$, there is some invariance property in the probabilistic evaluation given by (3) with respect to the initial choice of the measure M. More precisely, a set of interestingness measures can be divided into stable classes associated with different probability scales, respectively. For example, relative to the collection of 21 measures given in [27] and for the no relation random model denoted below by $\mathcal{N}_1$ (see subsection 2.1), the following interestingness measures lead to exactly the same likelihood of the link probabilistic measure: Support, Confidence, Interest, Cosine, Piatetsky-Shapiro, Jaccard, Kappa and ϕ coefficient.

Consequently our starting point in building the likelihood of the link probabilistic measure will be one of the respective entries of the 2×2 contingency table considered above: $n(a \wedge b)$, $n(a \wedge \neg b)$, $n(\neg a \wedge b)$ or $n(\neg a \wedge \neg b)$. With a clear intuitive sense it suffices to consider $n(a \wedge b)$ for the symmetrical association and $n(a \wedge \neg b)$ for the asymmetrical case. Remind that for the latter we have to set up $(a \rightarrow b)$ more strongly than $(b \rightarrow a)$.

Indeed, in the statistical literature, the symmetrical equivalence case ($a \rightarrow b$ and $b \rightarrow a$) has preceded the asymmetrical implicative one ($a \rightarrow b$). And in fact, several proposed indices in the data mining literature have a perfect symmetrical nature, that is to say that their expressions are invariant by substituting the ordered boolean pair (a, b) for (b, a). For example, relative to the above mentioned measures one may cite: Support, Interest, Cosine, Jaccard, Piatetsky-Shapiro, Kappa and ϕ.

The local context free form of the likelihood of the link probabilistic indices have been established first. The symmetrical version [19, 20, 18] has preceded the asymmetrical one [9, 23, 11]. The associated probabilistic scale was able to reveal fine structural relations on the attribute set $\mathcal{A}$, only when the number n of objects is lower than 10^3. But nowadays it is often necessary to work with large data (for example n greater than 10^6). And, in such situation, the latter scale becomes not enough discriminant in order to distinguish between different high values of the computed indices.

The "in the context" random model integrates in its construction the previous local model. But the probability scale becomes finely discriminant for any magnitude of n. This global model has been extensively validated theoretically [17, 7] and experimentally [20] in the framework of our hierarchical classification *LLA* (Likelihood of the Link Analysis) method (*Classification Ascendante Hiérarchique par Analyse de la Vraisemblance des Liens*) [20, 18, 21]. This validation aspect represents one important contribution of the analysis presented in this paper.

This analysis leads us to set up new absolute measures expressed in terms of the above relative frequencies $p(a \wedge b)$, $p(a \wedge \neg b)$, $p(\neg a \wedge b)$ and $p(\neg a \wedge \neg b)$ (see subsections 2.2 and 4.1). These measures appear as components of the χ^2 statistic.

Therefore, we begin in the second section by describing the probabilistic construction of the likelihood of the link index in the symmetrical comparison case and for the "context free" random model. The "in the context" random model will be expressed in section 3. Section 4 is devoted to implicative similarity which directly reflects the asymmetric nature of an association rule notion. Analysis of classical indices is performed in this section. Otherwise, local ("context free") and global ("in the context") interpretations are developed. In this section we will present the Probabilistic Normalized Index which enables to discriminate comparisons between association rules. Precisely, section 5 is devoted to experimental results validating the behaviour of this new index with respect to that locally built. We end in section 6 with a general conclusion giving the benefits and the prospects of this work.

2 "Context free" comparison between two boolean attributes

2.1 Building no relation (independence) hypothesis

Let (a, b) be a pair of boolean attributes provided from $\mathcal{A} \times \mathcal{A}$. We have introduced (see above) the set theoretic representation for this pair. We have also defined the cardinal parameters $n(a \wedge b)$, $n(a \wedge \neg b)$, $n(\neg a \wedge b)$ and $n(\neg a \wedge \neg b)$. Without loss of generality assume the inequality $n(a) < n(b)$.

Two distinctive but related problems have to be considered in comparing two boolean attributes a and b. The first consists of evaluating the degree of symmetrical equivalence relation $(a \leftrightarrow b)$. The second concerns asymmetrical implicative relation $(a \rightarrow b)$ called "association rule". Statistical literature has mainly focused the symmetrical association case. Many coefficients have been proposed for pairwise comparison of a set $\mathcal{A}$ of boolean attributes. All of them can be expressed as functions of the parameters $(n(a \wedge b), n(a), n(b), n)$. Most of them can be reduced to functions of the relative frequencies $(p(a \wedge b), p(a), p(b))$ associated with the absolute frequencies $n(a \wedge b)$, $n(a)$ and $n(b)$, relative to n. Thus, the parameter $n(a \wedge b)$ representing the number of objects where the conjunction $a \wedge b$ is true, appears as a fundamental basis of an association coefficient construction. We call this index a "raw" association coefficient. As a matter of fact, each of the association coefficients proposed in the literarture corresponds to a type of normalization of this raw index, with respect to the sizes $n(a)$ and $n(b)$. Indeed, tendency to a high or low values of $n(a \wedge b)$ is associated with high or low values of both parameters, respectively.

In these conditions, the first step of the normalization process we have adopted consists in introducing a probabilistic model of independence (or no relation) defined as a correspondence:

$$(\mathcal{O}(a), \mathcal{O}(b), \mathcal{O}) \rightarrow (\mathcal{X}, \mathcal{Y}, \Omega) \tag{4}$$

Three versions of this general model that we designate by $\mathcal{N}$, are considered. They lead to three distinct analytical forms of an association coefficient. Ω is associated with the object set $\mathcal{O}$, exactly ($\Omega = \mathcal{O}$) or randomly defined. For a given Ω, $\mathcal{X}$ and $\mathcal{Y}$ are defined by two independent random subsets of Ω, associated with $\mathcal{O}(a)$ and $\mathcal{O}(b)$, repectively. More precisely, the random model is built in such a way that Ω, $\mathcal{X}$ and $\mathcal{Y}$ respect exactly or on average the cardinalities n, $n(a)$ and $n(b)$, respectively. The random subsets $\mathcal{X}$ and $\mathcal{Y}$ can also be denoted by $\mathcal{O}(a^*)$ and $\mathcal{O}(b^*)$ where a^* and b^* are two independent random attributes associated with a and b, respectively.

Denoting by

$$s = n(a \wedge b) = card(\mathcal{O}(a) \cap \mathcal{O}(b)) \tag{5}$$

the above-mentioned raw index, the random raw index is defined by:

$$\mathcal{S} = n(a^* \wedge b^*) = card(\mathcal{X} \cap \mathcal{Y}) \tag{6}$$

The first form of normalization is obtained by standardizing s with respect to the probability distribution of S:

$$q(a,b) = \frac{s - \mathcal{E}(\mathcal{S})}{\sqrt{var(\mathcal{S})}} \tag{7}$$

where $\mathcal{E}(\mathcal{S})$ and $var(\mathcal{S})$ designate the mathematical expectation and the variance of $\mathcal{S}$.

By using normal distribution for the probability law of $\mathcal{S}$, this coefficient leads to the probabilistic index:

$$\mathcal{I}(a,b) = Pr\{\mathcal{S} \leq s|\mathcal{N}\} = Pr\{q(a^*,b^*) \leq q(a,b)|\mathcal{N}\} \tag{8}$$

that we call "local" likelihood of the link association index. Remind that "local" refers to the logical independence of its construction relative to the attribute set $\mathcal{A}$ from which a and b are taken.

For this index the similarity between a and b is measured by a probability value stating how much improbable is the bigness of the observed value of the raw index s. This probability is defined and computed under the independence hypothesis $\mathcal{N}$. Clearly, the index (9) is nothing but the complement to 1 of a $P - value$ in the sense of statistical hypotheses. However, its meaning does not refer to a conditional test [4] but to a conditional probabilistic evaluation, when the sizes $n(a)$ and $n(b)$ are given.

Three fundamental forms of the no relation (independence) hypothesis $\mathcal{N}$ have been set up [23, 20]. Let us denote them $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$. These are distinguished in their ways of associating with a given subset $\mathcal{O}(c)$ of $\mathcal{O}$, a random subset $\mathcal{L}$ of an Ω set corresponding to $\mathcal{O}$. Let us designate by $P(\Omega)$ the set of all subsets of Ω organized into levels by the set inclusion relation. A given level is composed by all subsets having the same cardinality.

Let us now make clearer the hypotheses $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$.

For $\mathcal{N}_1$, $\Omega = \mathcal{O}$ and $\mathcal{L}$ is a random element from the $n(c)$ level of $P(\Omega)$, provided by an uniform probability distribution. Then $\mathcal{L}$ is a random subset of $\mathcal{O}$ of size $n(c)$.

For $\mathcal{N}_2$, $\Omega = \mathcal{O}$. But the random model includes two steps. The first consists of randomly choosing a level of $P(\mathcal{O})$. Then $\mathcal{L}$ is defined as a random element of the concerned level, provided by an uniform distribution. More precisely, the level choice follows the binomial distribution with n and $p(c) = \frac{n(c)}{n}$ as parameters. Under these conditions, the probability of the k^{th} level, $1 \leq k \leq n$ is given by $C_n^k p(c)^k p(\neg c)^{n-k}$, where $p(\neg c) = \frac{n(\neg c)}{n}$.

$\mathcal{N}_3$ is defined by a random model with three steps. The first consists of associating with the object set $\mathcal{O}$ a random object set Ω. The only requirement for Ω concerns its cardinality N which is supposed following a Poisson probability law, its parameter being $n = card(\mathcal{O})$. The two following steps are similar to those of the random model $\mathcal{N}_2$. More precisely, for $N = m$ and an object set sized by m, $\mathcal{L}$ is a random part of Ω_0. $\mathcal{L}$ is defined only for $m \geq n(c)$ and in this case we define $\gamma = \frac{n(c)}{m}$. In these conditions, the probability to choose the level k of $\mathcal{P}(\Omega_0)$ is defined by the binomial probability $C_m^k \gamma^k (1-\gamma)^{n-k}$. And for a given level, the random choice of $\mathcal{L}$ is done uniformly on this level.

We established [23, 20] that the distribution of the random raw index $\mathcal{S}$ is:

- hypergeometric of parameters $(n, n(a), n(b))$, under the model $\mathcal{N}_1$;
- binomial of parameters $(n, p(a) * p(b))$, under the model $\mathcal{N}_2$;
- of Poisson of parameters $(n, n * p(a) * p(b))$, under the model $\mathcal{N}_3$.

2.2 The different versions of a statistically standardized index

The normalized form of the raw index s, according to equation (8) for the random models $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$, respectively, leads to the following indices [23]:

$$q_1(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b) * p(\neg a) * p(\neg b)}} \tag{9}$$

$$q_2(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b) * [1 - p(a) * p(b)]}} \tag{10}$$

and

$$q_3(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b)}} \tag{11}$$

Notice the perfect symmetry of $q_1(a,b)$ according to the following meaning:

$$q_1(a,b) = q_1(\neg a, \neg b) \tag{12}$$

As mentioned above (see introduction) we assume the positive form of the boolean attributes established in such a way that the proportional frequency of the "true" value is less than 0.5. Under this condition we have the following inequalities:

$$q_2(a,b) > q_2(\neg a, \neg b) \tag{13}$$

and

$$q_3(a,b) > q_3(\neg a, \neg b) \tag{14}$$

The last inequality is clearly more differentiated than the previous one. Therefore, we only consider the two most differentiated indices $q_1(a,b)$ and $q_3(a,b)$. One more reason for distinguishing these two indices concerns both formal and statistical aspects. Indeed, by considering the χ^2 statistic associated with the 2×2 contingency table crossing $(a, \neg a)$ and $(b, \neg b)$, we obtain:

$$\begin{aligned} \chi^2\{(a,\neg a),(b,\neg b)\} &= [q_1(a,b)]^2 \\ &= [q_3(a,b)]^2 + [q_3(a,\neg b)]^2 + \\ &\quad [q_3(\neg a,b)]^2 + [q_3(\neg a,\neg b)]^2 \end{aligned} \tag{15}$$

Thus, $q_3(a,b)$ defines the direct contribution of the entry (a,b) to the χ^2 statistic.

Dividing by $\sqrt{n}$ the indices q_1 and q_3, one obtains the respective associated indices γ_1 and γ_3. Correlative interpretation of these can be provided. Both are comprised between -1 and $+1$. But depending on $p(a)$ and $p(b)$, γ_3 is included into a more narrow interval than that of γ_1:

$$\gamma_1(a,b) = \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b) * p(\neg a) * p(\neg b)}} \tag{16}$$

and

$$\gamma_3(a,b) = \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b)}} \tag{17}$$

Now, let us designate by d_{ab} the density of the joint empirical probability with respect to the product of marginal probabilities, namely:

$$d_{ab} = \frac{p(a \wedge b)}{p(a) * p(b)} \tag{18}$$

The latter index is directly related to that t^{ab} introduced in [4] by means of the equation:

$$t^{ab} = d_{ab} - 1 \tag{19}$$

Then, $\gamma_1(a,b)$ and $\gamma_3(a,b)$ can be expressed in terms of d_{ab} and the marginal proportional frequencies. More precisely, we have:

$$\gamma_1(a,b) = \sqrt{\frac{p(a) * p(b)}{p(\neg a) * p(\neg b)}} * (d_{ab} - 1) \tag{20}$$

$$\gamma_3(a,b) = \sqrt{p(a) * p(b)} * (d_{ab} - 1) \tag{21}$$

2.3 Behaviour of the likelihood of the link local probabilistic index

The general expression of this index is given in (8) where the no relation (independence) hypothesis $\mathcal{N}$ is not yet specified. By substituting $\mathcal{N}$ with $\mathcal{N}_1$, $\mathcal{N}_2$ or $\mathcal{N}_3$, one has to replace q by q_1, q_2 or q_3, respectively. Even for n relatively high, an exact evaluation of the probability $Pr(\mathcal{S} \leq s|\mathcal{N}_i)(i = 1, 2$ or

3) can be obtained by means of a computer program. Cumulative distribution functions of hypergeometric, binomial or of Poisson laws are refered to the subscipt value i, $i = 1, 2$ ou 3, respectively. Nevertheless, for n large enough (for example greater than 100) and $p(a) * p(b)$ not too small, the probability law of $\mathcal{S}$ can be very accurately approximated by the normal distribution:

$$\mathcal{I}_i(a,b) = Pr\{\mathcal{S} \leq s|\mathcal{N}_i\} = \Phi[q_i(a,b)] = \Phi[\sqrt{n}\gamma_i(a,b)] \tag{22}$$

where Φ denotes the standardized cumulative normal distribution function and where $i = 1, 2$ or 3.

Consequently, if d_{ab} is clearly greater than unity and for n large enough, the local probabilistic index becomes very close to 1. On the contrary, if d_{ab} is lower than 1 and for n enough large, this index tends to 0. Thus, for n large enough and whatever the computing accuracy reached, this local probabilistic index is only able to discriminate on an attribute set $\mathcal{A}$ two classes of attribute couples: the positively and the negatively related.

Now, as it is usual in statistical inference, let us imagine the object set $\mathcal{O}$ obtained by means of a random sampling in a universe $\mathcal{U}$ of objects. Then, let us denote at the level of $\mathcal{U}$, by $\pi(a)$, $\pi(b)$ and $\pi(a \wedge b)$ the object proportional frequencies where the boolean attributes a, b and $a \wedge b$ have the "true" value. Note that $\pi(a)$, $\pi(b)$ and $\pi(a \wedge b)$ can be interpreted as probabilities associated with a "true" value for a random object taken with equiprobability distribution from $\mathcal{U}$. In these conditions let us designate by $\rho_i(a,b)$ the mathematical expression corresponding to $\gamma_i(a,b)$ and defined at the level of $\mathcal{U}$, $i = 1, 2$ or 3, respectively. In this fashion, $\gamma_i(a,b)$ defines an estimation of $\rho_i(a,b)$, whose accuracy is an increasing function of n [17].

Now, let us indicate by $\delta(a,b)$ the expression associated with (18), but at the level of $\mathcal{U}$. Clearly, we have the following properties for n enough large:

- $\delta(a,b) < 1$, $\mathcal{I}_i(a,b)$ [*cf.* (22)] tends to 0 ;
- $\delta(a,b) > 1$, $\mathcal{I}_i(a,b)$ tends to 1 ;

but, for $\delta(a,b) = 1$, $\mathcal{I}_i(a,b)$ can be considered as an observed value of an uniformly distributed random variable on the $[0,1]$ interval.

However, our statistical framework is restricted to the object set $\mathcal{O}$ and it is in this context that we have to achieve a probabilistic discriminant index. As shown above, this index cannot be obtained if we only have to compare in an absolute manner two boolean attributes a and b independently of the attribute set $\mathcal{A}$ from which they come. And indeed, if our universe is limited to this single couple of attributes $\{a,b\}$, the above proposed index $\mathcal{I}_i(a,b), \forall i = 1, 2$ ou 3, is sufficient. In fact, the objective consists of mutual comparison of many attribute pairs and generally of the whole set $P_2(\mathcal{A})$ of attribute pairs.

3 "In the context" comparison between two attributes

The context is determined by the set of attribute pairs of a set $\mathcal{A}$ of p boolean attributes:

$$\mathcal{A} = \{a^j | 1 \leq j \leq p\} \tag{23}$$

In this context, a probabilistic similarity between two boolean attributes will be proposed. This similarity will have a relative meaning with respect to the context. Retain from $\mathcal{A} \times \mathcal{A}$ the following cardinal structure:

$$\{(n(a^j \wedge a^k), n(a^j \wedge \neg a^k), n(\neg a^j \wedge a^k), n(\neg a^j \wedge \neg a^k)) | 1 \leq j < k \leq p\} \tag{24}$$

Also denote

$$\{n(a^j) | 1 \leq j \leq p\} \tag{25}$$

the sequence of the cardinalities of the subsets $\mathcal{O}(a^j), 1 \leq j \leq p$. At the same time, introduce the associated sequence of the proportional frequencies, relative to the total number of objects n. Thus the mathematical tables (24) and (25) give:

$$\{(p(a^j \wedge a^k), p(a^j \wedge \neg a^k), p(\neg a^j \wedge a^k), p(\neg a^j \wedge \neg a^k)) | 1 \leq j < k \leq p\} \tag{26}$$

and

$$\{p(a^j) | 1 \leq j \leq p\} \tag{27}$$

Now, reconsider the index $q_3(a, b)$ [*cf.* (11)] locally built by a centring and reducing process with respect to the no relation hypothesis $\mathcal{N}_3$. We have emphasized above the interesting asymmetrical property of this index where the similarity between rare attributes is clearly stressed [*cf.* (13)]. Precisely, we assume that the set of the p boolean attributes is established in such a way that:

$$n(a^j) \leq n(\neg a^j), 1 \leq j \leq p \tag{28}$$

As expressed in the introduction, a system of values such as (26) cannot be evaluated in the same way relatively to the induced equivalence or implicative relations, for any magnitude of the number n of objects. And that matters from two points of view: statistical and semantical. Moreover, the associations rules should be situated in a relative way.

To answer these two requirements let us first reconsider $q_3(a,b)$ [*cf.* (11)]. Now, in order to compare in a mutual and relative manner the set of attribute pairs, introduce the empirical variance of the index q_3 on the set $P_2(\mathcal{A})$ of two element parts of $\mathcal{A}$. This variance can be written:

$$var_e(q_3) = \frac{2}{p*(p-1)} \sum \{[q_3(a^j,a^k) - moy_e(q_3)]^2 | 1 \leq j < k \leq p\} \quad (29)$$

where

$$moy_e(q_3) = \frac{2}{p*(p-1)} \sum \{q_3(a^j,a^k) | 1 \leq j < k \leq p\} \quad (30)$$

defines the mean of the index q_3 on $P_2(\mathcal{A})$.

For relative comparison between two attributes belonging to $\mathcal{A}$, we introduce the globally normalized index q_3. For example, consider the comparison between two given attributes a^1 and a^2. The new index $q_3^g(a^1,a^2)$ is defined as follows:

$$q_3^g(a^1,a^2) = \frac{q_3(a^1,a^2) - moy_e(q_3)}{\sqrt{var_e(q_3)}} \quad (31)$$

This index corresponds to the relative and directed contribution of $q_3(a^1,a^2)$ to the empirical variance $var_e(q_3)$.

Under these conditions, the likelihood of the link probabilistic index is conceived with respect to a global independence hypothesis where we associate with the attribute set $\mathcal{A}$ [*cf.* (23)] a random attribute set:

$$\mathcal{A}^* = \{a^{j*} | 1 \leq j \leq p\} \quad (32)$$

where the different attributes are mutually independent with respect to the no relation hypothesis $\mathcal{N}_3$. This index can be written:

$$P_g(a^1,a^2) = Pr\{q_3^g(a^{1*},a^{2*}) \leq q_3^g(a^1,a^2) | \mathcal{N}_3\} \quad (33)$$

where

$$q_3^g(a^{1*},a^{2*}) = \frac{q_3(a^{1*},a^{2*}) - moy_e(q_3^*)}{\sqrt{var_e(q_3^*)}} \quad (34)$$

Theoretical and experimental proofs [17, 7] show that the probabilistic index can be computed by means of the following equation :

$$P_g(a^1,a^2) = \Phi[q_3^g(a^1,a^2)] \quad (35)$$

where Φ denotes the standardized normal cumulative distribution.

Let us point out that the table of probabilistic similarity indices

$$\{P_g(a^j, a^k)|1 \leq j < k \leq p\} \tag{36}$$

is that taken into account in the LLA ascendant hierarchical classification method [18, 21]. Otherwise and for any data analysis method working with dissimilarities, our approach can provide the following table of dissimilarity indices:

$$\{\mathcal{D}(j,k) = -Log_2[P_g(a^j, a^k)]|1 \leq j < k \leq p\} \tag{37}$$

that we call " informational dissimilarity table ".

4 Implicative similarity index

4.1 Indices independent of n and of the context

So far we have been interested in the evaluation of the symmetrical equivalence relation degree between two boolean attributes a and b belonging to a set $\mathcal{A}$ of boolean attributes [*cf.* (23)]. For this purpose we refered to independence statistical hypothesis as a basis to establish an adequate measure. The built indices are clearly situated with respect to this hypothesis. Now, we have to evaluate the asymmetrical implicative relation of the form $a \rightarrow b$. Such a relation is completely satisfied at the level of the object set $\mathcal{O}$ if the subset $\mathcal{O}(a)$ characterized by a true value of a is included in the subset $\mathcal{O}(b)$ characterized by a true value of b. In practice this event is very rare. When, without having strict inclusion, there are clear experimental situations defined at the level of the object set $\mathcal{O}$, where partial inclusion is more or less strong. And in such situations, we have to evaluate the tendency of b knowing a.

Let us reconsider briefly the complete inclusion situation observed at the level of the object set $\mathcal{O}$. It can be interesting to study it by introducing parametrization with $(n, p(a), p(b))$ where $p(a \wedge \neg b) = 0$. One can also [26] be interested in some aspects of the probability law of the conditional relative frequency $p(b^*|a^*)$ under a very specific model. For the latter it is assumed that both cardinalities are known : the size $N(a)$ of the $\mathcal{U}$ subset where a is true and the size $n(a)$ of the $\mathcal{O}$ subset where a is true.

Now, let us consider the most realistic and the most frequent case where the number of objects $n(a \wedge \neg b)$ is small without being null. In these conditions, we are interested in evaluating the relative smallness of the number of objects $n(a \wedge \neg b)$ where a is true and where b is false. These objects contradict the implicative relation $a \rightarrow b$. In order to evaluate this smallness, most of the proposed indices try to neutralize the size influence of $n(a)$ and $n(b)$. However notice that for $n(a)$ and $n(b)$ fixed, $n(a \wedge \neg b)$ "small", $n(a \wedge b)$ "large", $n(\neg a \wedge b)$ "small" and $n(\neg a, \neg b)$ "large" correspond to concordant phenomenons. Thus, some proposed indices for the asymmetrical implicative case have a perfect symmetrical nature with respect to the ordered pair (a, b). Most if not all of them can be situated with respect to the empirical independence hypothesis defined by $p(a \wedge b) = p(a) * p(b)$. According to [26] the easiest one is $(p(a \wedge b) - p(a) * p(b))$. The latter corresponds to the common numerator of $q_1(a, b)$, $q_2(a, b)$ and $q_3(a, b)$ (cf. 9,10,11). This index and $\gamma_1(a, b)$ are reported in [26] but at the level of the object universe $\mathcal{U}$. γ_1 is nothing else than the K. Pearson coefficient [25]. A symmetrical index called "interest measure" is also proposed in [6]. This index has been denoted by d_{ab} (cf. 18). It is directly related to the contribution of the entry (a, b) to the χ^2 statistic. This contribution can also be expressed by $[q_3(a, b)]^2$.

Nevertheless, many of the proposed indices for evaluating the strength of an implication $a \rightarrow b$ are asymmetrical. Clearly, they stress the smallness of $n(a \wedge \neg b)$ with respect to the bigness of $n(a \wedge b)$. Let us describe some of them. For coherent reasons in the following development but without explicit intervention, suppose the inequality $n(a) \leq n(b)$ which makes possible the total inclusion of $\mathcal{O}(a)$ in $\mathcal{O}(b)$.

The easiest and the most direct index that we have to mention is that called " the confidence index" [1]. It is defined by the conditional proportion $p(b|a) = \frac{p(a \wedge b)}{p(a)}$. It varies from 0 to 1. The 0 value is associated with disjunction between $\mathcal{O}(a)$ and $\mathcal{O}(b)$, when the 1 value is reached in case of inclusion of $\mathcal{O}(a)$ into $\mathcal{O}(b)$. An interesting comparison analysis of this basic index with different indices proposed in the literature is developed in [14]. The Loevinger index [24] is also a very classical and very known one. Respective to the introduced notations [*cf.* (18)] it can be defined by the following equation :

$$\mathcal{H}(a, b) = 1 - d(a, \neg b) \tag{38}$$

Let us now suppose the two "natural" inequalities $p(a) \leq p(b)$ and $p(a) \leq p(\neg b)$. Under these conditions, the index value varies from 1 in case of complete inclusion $\mathcal{O}(a) \subset \mathcal{O}(b)$, goes through the 0 value for the statistical independence and reachs the negative value $-[\frac{p(b)}{p(\neg b)}]$ in case where $\mathcal{O}(a) \subset \mathcal{O}(\neg b)$; that is to say, where the opposite implication $a \rightarrow \neg b$ holds.

The Lœvinger index can also be written :

$$\mathcal{H}(a, b) = \frac{p(a \wedge b) - p(a) * p(b)}{p(a) * p(\neg b)} \tag{39}$$

Then, it corresponds to an asymmetrical reduction with respect to (a, b) of the first index proposed by G. Piatetsky-Shapiro [26].

One can also clearly situate the $\mathcal{H}(a,b)$ index with respect to $\gamma_3(a,\neg b)$. $\mathcal{H}(a,b)$ is obtained by reducing the centred index $[p(a\wedge\neg b)-p(a)*p(\neg b)]$ by means of $p(a)*p(\neg b)$ when the reduction is performed with $\sqrt{p(a)*p(\neg b)}$ in the $-\gamma_3(a,\neg b)$ index. Consequently, $-\gamma_3(a,\neg b)$ defines a new implicative index only depending on $[p(a\wedge b), p(a), p(b)]$. Obviously the 0 value characterizes statistical independence. On the other hand, the logical implication $a\rightarrow b$ is obtained for the value $\sqrt{p(a)*p(\neg b)}$. Otherwise, the index value can decrease till the value $-p(b)*\sqrt{\frac{p(a)}{p(\neg b)}}$.

Therefore and clearly, one can propose the following discriminant "free context" index :

$$-\gamma_3(a,\neg b)=\frac{-p(a\wedge\neg b)+p(a)*p(\neg b)}{\sqrt{p(a)*p(\neg b)}} \tag{40}$$

This index is exactly the opposite of the direct contribution of the entry $(a,\neg b)$ to the χ^2/n coefficient. The index $-\gamma_3(a,\neg b)$ appears coherent with the "in the context" construction process of the likelihood of the link probabilistic index (see the above section 3). It can be seen that the two extreme value limits are included in the interval $[-1,+1]$ when the lowest negative boundary of $\mathcal{H}(a,b)$ can potentially reach any negative value. More precisely and under the inequalities $(p(a)\leq p(b)$ and $p(a)\leq p(\neg b))$, we establish that the two boundaries are comprised in the interval $[-p(b), p(\neg b)]$.

Now, by considering the coherence condition $p(a)\leq p(b)$ which enables the complete inclusion $\mathcal{O}(a)\subset\mathcal{O}(b)$ and, as said in the introduction, by assuming the significant conditions $p(a)\leq p(\neg a)$ and $p(b)\leq p(\neg b)$ one can establish that the minimal and maximal values of $\gamma_3(a,\neg b)$ are -0.5 and $+0.5$, respectively. The above limit $p(\neg b)$ is then reduced. Consequently and under the mentioned coherence conditions, we can obtain an index whose value ranges from 0 to 1 by setting :

$$\eta_3(a,b)=0.5-\gamma_3(a,\neg b) \tag{41}$$

Otherwise and under the above coherent conditions, the minimal boundary for the index $\mathcal{H}(a,b)$ is greater or equal to -1. Then an index denoted by $\mathcal{K}(a,b)$, deduced from $\mathcal{H}(a,b)$ and comprised between 0 and 1 can be put in the following form :

$$\mathcal{K}(a,b)=\frac{1}{2}(\mathcal{H}(a,b)+1) \tag{42}$$

Notice that the two new indices have the common value $\frac{1}{2}$ in case of statistical independence.

It has been shown that the presented above indices can be expressed in terms of the different components of the χ^2 statistic associated with the 2×2 contingency table defined by the crossing $\{a, \neg a\} \times \{b, \neg b\}$. We are now going to present two indices which employ the mutual information statistic associated with this contingency table. Three formal versions can be considered for this statistic :

$$\begin{aligned}
\mathcal{E} &= p(a \wedge b)\log_2(d(a,b)) + p(a \wedge \neg b)\log_2(d(a,\neg b)) \\
&+ p(\neg a \wedge b)\log_2(d(\neg a,b)) + p(\neg a \wedge \neg b)\log_2(d(\neg a,\neg b)) && (43)\\
&= E(a) - p(b)E(a|b) - p(\neg b)E(a|\neg b) && (44)\\
&= E(b) - p(a)E(b|a) - p(\neg a)E(b|\neg a) && (45)
\end{aligned}$$

where $E(x)$ denotes the entropy of the binary distribution $(p(x), p(\neg x))$ and where $E(x|y)$ denotes that of the conditional distribution $(p(x|y), p(\neg x|y))$, x and y, indicating two boolean attributes.

Precisely, the Goodman & Smith [8] J-measure corresponds to the sum of the first two terms of the previous first equation (43). In these, the boolean attribute a is taken positively, when the negated attribute $\neg a$ is considered in the sum of the last two terms of (43). A second index proposed by R. Gras [10] has fundamentally an entropic conception. It is called "inclusion" index and it takes the following form :

$$\tau(a,b) = \sqrt{G(b|a) * G(\neg a|\neg b)} \qquad (46)$$

where $G(x|y)$ is defined by the square root of

$$G^2(x|y) = \begin{cases} 1 - E^2(x|y) \text{ if } p(\neg x \wedge y) \leq \frac{1}{2} * p(y) \\ 0 \text{ if not} \end{cases} \qquad (47)$$

This index employs the conditional entropies $E(b|a)$ and $E(\neg a|\neg b)$ which are associated with the binary distributions $(p(b|a), p(\neg b|a))$ and $(p(\neg a|\neg b), p(a|\neg b))$, respectively. The former entropy can be obtained as a constitutive component of the equation (45) expressing the mutual information $\mathcal{E}$, when the latter entropy is defined as a component element of the preceding equation (44).

Notice that a high value of the index $1 - E^2(b|a)$ expresses two opposite inclusion tendencies. The first one consists of $\mathcal{O}(a) \subset \mathcal{O}(b)$ and the other logical opposite $\mathcal{O}(a) \subset \mathcal{O}(\neg b)$ $(E(b|a) = E(\neg b|a))$. Similarly, a high value of $1 - E^2(\neg a|\neg b)$, reflects two opposite inclusion tendencies corresponding to $\mathcal{O}(\neg b) \subset \mathcal{O}(\neg a)$ and $\mathcal{O}(\neg b) \subset \mathcal{O}(a)$ $(E(\neg a|\neg b) = E(a|\neg b))$. However, taking into account the condition included in equation (47), a strictly positive value of $\tau(a,b)$ is constrained by $p(b|a) > p(\neg b|a)$. Consequently, a strictly positive value of the inclusion index cannot occur if one of both conditional probabilities is lower than 0.5. Now, there may be situations where $p(b|a)$ is high enough (clearly greater than 0.5) and where $p(\neg a|\neg b)$ is low enough (notably lower than 0.5). And in these, there is no reason to reject *a priori* an implicative tendency value for $a \to b$. This weakness of the inclusion index is somewhat balanced by its quality consisting of taking into account both implicative evaluations : $a \to b$ and $\neg b \to \neg a$.

4.2 Comparing local implicative of the link likelihood and entropic intensity indices

Conceptually, for associating symmetrically two boolean attributes a and b, the likelihood of the link probabilistic approach evaluates under a random model of no relation how much unlikely is in probability terms the relative bigness of $n(a \wedge b)$. The respective influences of the $n(a)$ and $n(b)$ sizes are neutralized in this model. This idea has been extensively developed in the framework of the LLA ascendant hierarchical classification of descriptive attributes [20]. It has been adapted by R. Gras [9], [23] in the asymmetrical implicative case. In the latter, one has to evaluate how much is unlikely the smallnes of $n(a \wedge \neg b)$ with respect to a random no relation model $\mathcal{N}$, neutralizing the respective influences of $n(a)$ and $n(b)$. The index can be written :

$$\begin{aligned} \mathcal{I}(a,b) &= 1 - Pr\{n(a^* \wedge \neg b^*) < n(a \wedge \neg b)|\mathcal{N}\} \\ &= Pr\{n(a^* \wedge \neg b^*) \geq n(a \wedge \neg b)|\mathcal{N}\} \end{aligned} \tag{48}$$

where (a^*, b^*) denotes the random ordered attribute pair associated with (a, b) under a random model $\mathcal{N}$ defining an independence hypothesis.

Let us designate by u the index $n(a \wedge \neg b)$ and by u^*, the random associated index $n(a^* \wedge \neg b^*)$ under the hypothesis of no relation $\mathcal{N}$. Then, the standardized index (by centring and reducing u) takes the following form :

$$q(a, \neg b) = \frac{u - \mathcal{E}(u^*)}{\sqrt{var(u^*)}} \tag{49}$$

where $\mathcal{E}(u^*)$ and $var(u^*)$ denote the mathematical expectation and the variance of u^*, respectively.

As expressed above three versions of the random model $\mathcal{N}$ have been set up : $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$. These lead for the random index u^*, to hypergeometric, binomial and of Poisson probability laws, respectively. $\mathcal{N}_1$ and $\mathcal{N}_3$ are the most differentiated models [23]. By designating $q_i(a, \neg b)$, the index u standardized with respect $\mathcal{N}_i$, we have :

$$q_1(a, \neg b) = q_1(\neg a, b) = -q_1(a, b) = -q_1(\neg a, \neg b) \tag{50}$$

Remark that for the random model $\mathcal{N}_1$, the implicative form of the index is exactly equivalent to the symmetrical case. For $\mathcal{N}_3$, we have with the condition $p(a) < p(b)$:

$$|q_3(a, \neg b)| > |q_3(b, \neg a)| \tag{51}$$

A natural condition in order to consider the evaluation of the association rule $a \to b$ can be defined by a negative value of $n(a \wedge \neg b) - (\frac{n(a)*n(\neg b)}{n})$. This expression represents the numerator of $q_3(a, \neg b)$. It is identical to that of $n(b \wedge \neg a) - (\frac{n(b)*n(\neg a)}{n})$ associated with the index $q_3(b, \neg a)$ corresponding to the opposite implication $b \to a$. However, since $n(b) > n(a)$, the latter is more difficult to accept. The inequality (51) consists of a coherent statement since, for the local likelihood of the link index associated with $\mathcal{N}_3$, we have (see (48)) :

$$\mathcal{J}_3(a, b) > \mathcal{J}_3(b, a) \tag{52}$$

To be convinced of this property, consider the excellent normal approximation for n enough large, of the probability Poisson law of $n(a^* \wedge \neg b^*)$ and $n(b^* \wedge \neg a^*)$, respectively, under the $\mathcal{N}_3$ model :

$$\mathcal{J}_3(a, b) = 1 - \Phi(q_3(a, \neg b)) \tag{53}$$

and

$$\mathcal{J}_3(b, a) = 1 - \Phi(q_3(b, \neg a)) \tag{54}$$

where Φ is the cumulative distribution function of the standardized normal law.

Moreover, the necessary and sufficient condition to have (52) is $n(a) < n(b)$. Mostly we have $p(a) < p(b) < \frac{1}{2}$. With this condition one obtains the following inequalities which comprise that (52) :

$$\mathcal{I}_3(\neg a, \neg b) < \mathcal{J}_3(b, a) < \mathcal{J}_3(a, b) < \mathcal{I}_3(a, b) \tag{55}$$

where $\mathcal{I}_3$ indicates the local likelihood of the link probabilistic index defined in (22).

Now, consider two situations that we denote by I and II where $\mathcal{O}(a)$ and $\mathcal{O}(b)$ have the same relative position. We mean that the proportional frequencies induced by $n(a \wedge b)$, $n(a \wedge \neg b)$ and $n(\neg a \wedge b)$ remain constant. But we suppose variation for the relative frequency induced by $n(\neg a \wedge \neg b)$ between I and II. Notice that every similarity index symmetrical or asymmetrical based on relative proportions defined into $\mathcal{O}(a \vee b)$ cannot distinguish I and

II, see for example the famous Jaccard index [13] or the confidence index $(a \to b) = \frac{n(a \wedge b)}{n(a)}$. This statement becomes false in case of an association rule whose conception depends logically on $\mathcal{O}(\neg a \wedge \neg b)$. In order to illustrate this point and to explain the behaviour of the local probabilistic index which is directly related to $q_3(a, \neg b)$, consider for $n = 4000$ the two following situations I and II. The situation I is described by $n(a \wedge b) = 200$, $n(a \wedge \neg b) = 400$ and $n(\neg a \wedge b) = 600$; when, the situation II relative to the attribute ordered pair that we denote by (a', b'), is characterized by $n(a' \wedge b') = 400$, $n(a' \wedge \neg b') = 800$ and $n(\neg a' \wedge b') = 1200$. Thus, $n(a' \wedge b')$, $n(a' \wedge \neg b')$ and $n(\neg a' \wedge \neg b')$ are obtained as twice $n(a \wedge b)$, $n(a \wedge \neg b)$ and $n(\neg a \wedge \neg b)$, respectively. In these conditions, it is not surprising to observe that the situation I corresponds to a strong implication $a \to b$ according to the q_3 value, $q_3(a, \neg b) = -3.65$; when, for the situation II, the implication $a' \to b'$ vanishes, $(q_3(a', \neg b') = 2.98)$. In fact, by comparing the respective sizes of $\mathcal{O}(a)$ and $\mathcal{O}(b)$ for one hand and, $\mathcal{O}(a')$ and $\mathcal{O}(b')$ for the other hand, we can realize that the inclusion degree of $\mathcal{O}(a)$ into $\mathcal{O}(b)$ is much more exceptional than that of $\mathcal{O}(a')$ into $\mathcal{O}(b')$. This latter should reach the value 578 for $n(a' \wedge b')$ to have $q_3(a', \neg b') = q_3(a, \neg b) = -3.65$. This phenomenon is amplified when we multiply all the cardinalities by a same coefficient greater than 1. Thus, with a multiplicative factor 10 one obtains $q_3(a, \neg b) = -11.55$ et $q_3(a', \neg b') = 9.43$. Moreover, with a multiplicative factor equal to 100 one obtains $q_3(a, \neg b) = -36.51$ and $q_3(a', \neg b') = 29.81$.

Consequently, for n enough large the local probabilistic index $\mathcal{J}_3(a, b)$ looses its discriminant power for pairwise comparing many implication rules. The solution proposed by [10] consists of combining a geometric mean the initial index $\mathcal{J}_3(a, b)$ (denoted by φ) with the inclusion index $\tau(a, b)$ [*cf.* (46)], in order to obtain the index called by R. Gras " entropic intensity :

$$\Psi(a, b) = \sqrt{\varphi(a, b) * \tau(a, b)} \tag{56}$$

But indices $\varphi(a, b)$ and $\tau(a, b)$ mixed into a unique one, are very different in their conceptions though they are related in an implicit way logically and statistically. This relation is difficult to analyze in spite of an established formal link between the χ^2 and the mutual information statistics [3]. In these conditions and if n is not too large in order to allow significant contributions of the two components of the entropic intensity, we cannot control the respective parts of these components in the index value $\Psi(a, b)$. Now, for n large enough and increasing, $\varphi(a, b)$ becomes closer and closer of 0 or 1 and then the index $\Psi(a, b)$ tends quickly to 0 or to $\sqrt{\tau(a, b)}$. Deeper formal and statistical analysis would be interesting to be concretely provided for this combined index in the framework of interesting experimental results.

The next subsection is devoted to our approach in which a pure probabilistic framework is maintained.

4.3 Implicative contextual similarity of the likelihood of the link

To build an implicative probabilistic index which remains discriminant whatever the value of n we proceed to a global reduction of the implicative similarities of the form $q_3(a, \neg b)$, with respect to an interesting set of ordered attribute pairs. This solution has been previously proposed, but with less consistency in [23]. More precisely, this normalization has to be performed in the context of a data base comprising attribute couples of which the corresponding association rules have to be compared mutually and in a relative manner. Under these conditions, we are placed into a dependency statistical structure and that could be more or less strong.

This method consists simply in transposing to the asymmetrical case the normalization adopted for the symmetrical one [*cf.* section 3]. Excellent experimental resuts have been obtained in practicing ascendant hierarchical classification according to the LLA likelihood of the link method [20, 18, 21].

A significant problem concerns choosing the set of attribute ordered pairs as a basis for normalization. In the following experimental scheme this basis is defined by all the distinct attribute couples of $\mathcal{A}$ [*cf.* 23]. We will denote the latter by its graph :

$$\mathcal{G}_0 = \{(j, k) | 1 \leq j \neq k \leq p\} \tag{57}$$

In [23] a selective choice have been recommended where the reference graph is

$$\mathcal{G}_1 = \{(j, k) | (1 \leq j \neq k \leq p) \wedge (n(a^j) < n(a^k))\} \tag{58}$$

so the full absorption of $\mathcal{O}(a^j)$ by $\mathcal{O}(a^k)$ can be achieved.

In the data mining community, an implication as $a \rightarrow b$ is taken into account only if indices such like $support(a \rightarrow b) = \frac{n(a \wedge b)}{n}$ and $confidence(a \rightarrow b) = \frac{n(a \wedge b)}{n(a)}$ are respectively higher than thresholds s_0 and c_0 defined by an expert [1, 12]. In this context, we will focus on the data base defined by the graph:

$$\begin{aligned}\mathcal{G}_2 = \{(j, k) | (1 \leq j \neq k \leq p) &\wedge (n(a^j) < n(a^k)) \\ &\wedge (support(a^j \rightarrow a^k) > s_0) \\ &\wedge (confidence(a^j \rightarrow a^k) > c_0)\}\end{aligned} \tag{59}$$

Now and with respect to a given graph $\mathcal{G}_i$ ($i = 0, 1$ or 2), global normalization leads us to replace the "local" index $q_3(a^j, a^k)$, by the "global" index $q_3^g(a^j, a^k)$ defined as follows :

$$q_3^g(a^j, a^k) = \frac{q_3(a^j, a^k) - mean_e\{q_3|\mathcal{G}_i\}}{\sqrt{var_e\{q_3|\mathcal{G}_i\}}} \tag{60}$$

where $mean_e\{q_3|\mathcal{G}_i\}$ and $var_e\{q_3|\mathcal{G}_i\}$ represent the empirical mean and variance of q_3 on $\mathcal{G}_i$, respectively.

The no relation or independence random model considered is that $\mathcal{N}_3$ (see above). But here this model has to be interpreted globally by associating with the whole attribute set $\mathcal{A}$ a set $\mathcal{A}^*$ of independent random attributes [*cf.* (32)]. In these conditions, $q_3^g(a^{j*}, a^{k*})$ follows a standardized normal law whose cumulative distribution function is, as above, denoted by Φ. Thus, in order to evaluate an association rule $(a^j \rightarrow a^k)$ taking its place in $\mathcal{G}_i$, the likelihood of the link probabilistic discriminant index is defined by the equation :

$$\mathcal{J}_n(a^j, a^k) = 1 - \Phi\left[q_3^g(a^j, a^k)\right] \tag{61}$$

5 Experimental results

5.1 Experimental scheme

Two kinds of experiments have been performed using the "mushrooms" database [5]. This database contains 8124 individuals described by 22 attributes and additionally an attribute class. The latter has been considered at the same level as the others attributes and all of them have been transformed into boolean attributes. By this way, the database is made of 125 attributes describing 8124 individuals. Our choice of this data base is motivated by studying the behaviour of the tested indices on real data and not on artificial ones.

The experimental process can be described as follows :

1. Compute all the couples (a, b) such that $n(a) < n(b)$
2. For each couple, compute $q_3(a, \neg b)$ and $\mathcal{J}(a, b)$ [*cf.* (48)]
3. Then compute q_3 mean and standard deviation for all the couples (a, b) and normalize the q_3 index before using the normal law.

In the first experiments, selection of couples (a, b) is only controlled by the following condition: $n(a) \neq 0$, $n(b) \neq 0$ and $n(a \wedge b) \neq 0$.
In the second experiments, some support and confidence constraints are used to control the selected couples. Remind that the support and the confidence of a couple (a, b) are equal to $Pr(a \wedge b)$ and $Pr(b|a) = \frac{Pr(a \wedge b)}{Pr(a)}$, respectively. All the selected couples in the second experiments, are such that $support(a, b) > s_0$ and $confidence(a, b) > c_0$.

These second experiments allow us to analyze the behaviour of several indices. The main goal of all these experiments is to observe the behaviour of the built indices when the size of $\mathcal{O}$ increases without any modification of the cardinalities of $\mathcal{O}(a)$, $\mathcal{O}(b)$ and $\mathcal{O}(a) \cup \mathcal{O}(b)$. In order to do that, a couple (a, b) is selected according to the constraints of the experimental scheme. Then, the value of n is increased without modifying the values of $n(a)$, $n(b)$ and $n(a \wedge b)$. This technique can be compared to adding occurences for which all the concerned attributes have a false value.

We used the following algorithm to realize our experiments:

(1) Select all the couples (a, b) verifying the experimental scheme constraints
(2) For each couple, compute $q_3(a, \neg b)$
(3) Compute mean and standard-deviation of all the q_3 values
(4) For each couple, compute the local likelihood of the link probabilistic index and the globally normalized one
(5) Increase $n : n \leftarrow n + constant$
(6) Repeat (2-5) until $n < threshold$

5.2 Detailed results

The first set of experiments (no constraints on the couples (a, b)) allow us to show many relevant properties for our normalized index.

- this index is discriminant whatever is the n value
- moreover, results show that the behaviour of our index is more ignificant for rules $a \rightarrow b$ when $n(a) < n(b)$

Figures 1 and 2 show that the normalized index has a discriminant behaviour unlike the local index (always equal to 1), for attributes a and b considered.

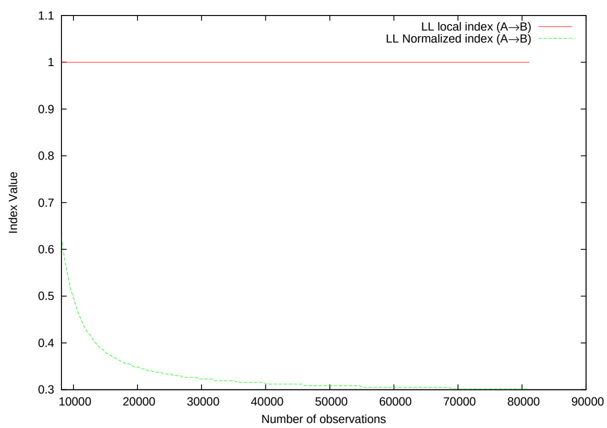

Fig. 1. $n(a) = 192$, $n(b) = 1202$, $n(a \wedge b) = 96$, $s_0 = 0$, $c_0 = 0$

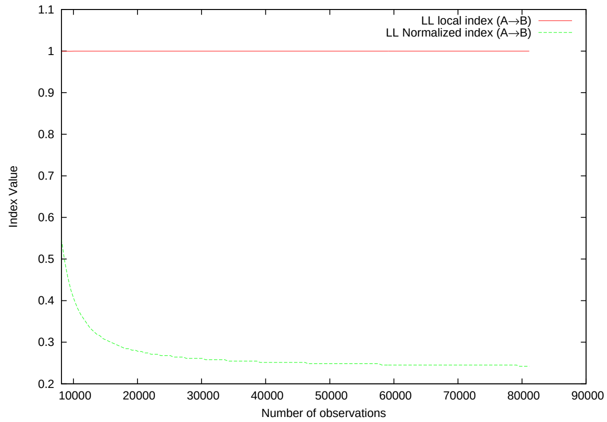

Fig. 2. $n(a) = 192$, $n(b) = 828$, $n(a \wedge b) = 64$, $s_0 = 0$, $c_0 = 0$

We can also show, on Figure 3, that the normalized index reachs low values, even though the local index tends to unity and this, when $n(a \wedge b)$ is small compared to $n(a)$ (see Figure 4 and 3).

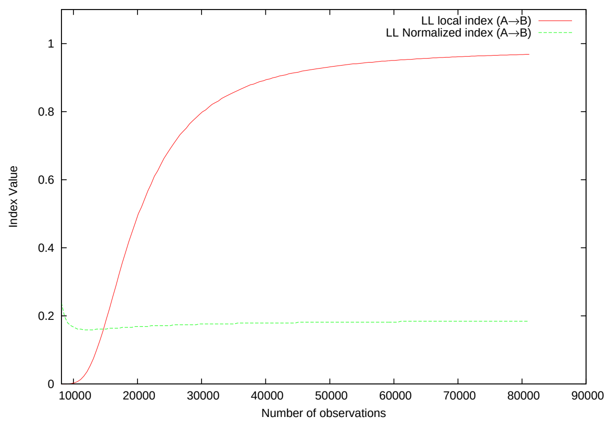

Fig. 3. $n(a) = 452$, $n(b) = 2320$, $n(a \wedge b) = 52$, $s_0 = 0$, $c_0 = 0$

5.3 Robust version of the normalized probabilistic index on a selected attribute couples

In the second series of experiments we focus on couples (a, b) whose support and confidence values are higher than thresholds defined by the user. For all of our experiments, we have used the thresholds $s_0 = 0.1$ and $c_0 = 0.9$. Then,

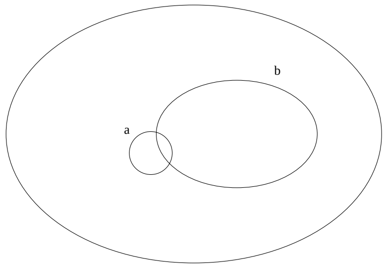

Fig. 4. $n(a) = 452$, $n(b) = 2320$, $n(a \wedge b) = 52$

the couples (a, b) under study will be such that $support(a \to b) > 0.1$ and $confidence(a \to b) > 0.9$. In fact these, thresholds are usually considered in rule extraction context [15, 2]. Furthermore, similar values have been used with the mushrooms database.

Achieved results show that, for this reduced set of attribute couples (a, b), our normalized index is always more discriminant than the local one. Moreover, as we can see on Figure 7, the normalized index is more discriminant in this experimental conditions than in the previous ones (see Figure 6). The shown curves are associated with the configuration presented in Figure 5.

In the second series of experiments, strong relationships characterize all the studied couples. In this case, with n increasing the relationship presented in Figure 5 becomes less and less significant relative to all the other relations with the same conditions. Therefore, the filtering step based on thresholds s_0 and c_0 allows us to only retain a set of relations having high values for all the considered indices.

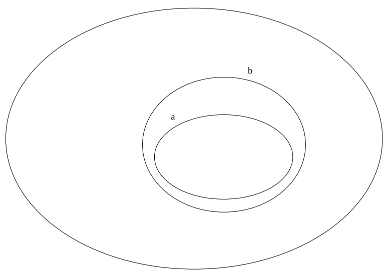

Fig. 5. $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$

Let us now consider the above situation (see Figure 5) where $\mathcal{O}(a) \subset \mathcal{O}(b)$. In this case, $n(a) = 1296$, $n(b) = 2304$ and $n(a \wedge b) = 1296$. Figure 6 shows the evolution of the normalized index when the number of objects increases from $n = 8124$ by adding fictive objects where all the attributes are false. The

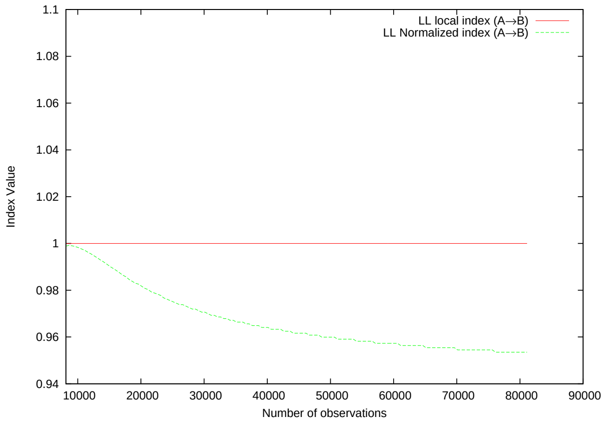

Fig. 6. $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0$, $c_0 = 0$

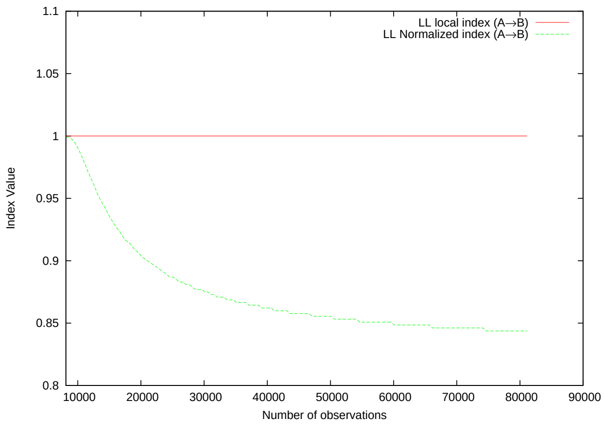

Fig. 7. $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0.1$, $c_0 = 0.9$

index value remains strong, higher than 0.98 as n reaching 80000 and starting from $n = 8124$. Nevertheless, this value decreases reaching a stable level when n increases. The reason is that when n increases the above implication becomes less and less distinctive with respect to the other implications. Such situation must concern cases where $n(a)$ and $n(b)$ are high and near each other.

Consider now the last series of experiments (Figure 7), it is not surprising to observe such small values for our normalized index when n increases. Indeed, this situation, described by the graph $\mathcal{G}_2$ (see 59) concerns attribute couples (a_j, a_k) where the implication relation is very strong. Thus for example by taking $n = 80000$, we have for every selected (a_j, a_k), $n(aj \wedge a_k) \geq 8000$ and $\frac{n(a_j \wedge a_k)}{n(a_j)} \geq 0.9$. But the indices are computed on the basis of the initial object set whose size is $n = 8214$. These indices must correspond to nearly inclusions of a very "big" $\mathcal{O}(a_j)$ into a slightly bigger $\mathcal{O}(a_k)$.

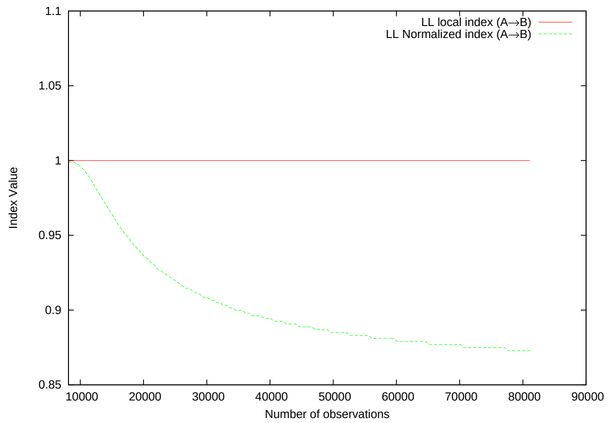

Fig. 8. $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0.1$, $c_0 = 0.5$

When the minimal confidence threshold decreases to 0.5, we observe, on Figure 8, that the relation considered becomes more significant with respect to the other relations. This behaviour is not surprising and can be explained by the fact that the set of studied relations includes relations with lower degree of confidence than in the case presented in Figure 7. As a consequence of all our analysis one can say that the new probabilistic normalized index by global reduction is discriminant and reflects in some way the statistical surprise of a rule in the context of other rules.

6 Conclusion

In order to evaluate an association rule $a \to b$, many indices have been proposed in data mining literature. Most of them have an absolute meaning and only depend on the concerned ordered pair (a, b) of attributes. This dependence is expressed in terms of proportional frequencies defined by the crossing between the two attributes on a learning set, $p(a)$, $p(b)$ and $p(a \wedge b)$. The formal aspects of these indices are analyzed and mutually compared. In their construction, the importance of empirical statistical independence is set up. Mainly, these indices are compared with our approach in defining probabilistic similarity measure associated with a notion of likelihood of the link with respect to independence hypothesis. Indeed, this notion is expected to capture that of "interestingness measure" setting up surprising rules. This similarity can be symmetrical translating equivalence relation degree or asymmetrical, reflecting asymmetrical implicative relation.

The first conception of the likelihood of the link similarity measure is local, i.e. only depends on the attribute pair to be compared. Unfortunately, this local version looses its discriminant power when the data size becomes large enough. And one of the major objectives of this paper consists in building a likelihood of the link probabilistic index associated in a specific relative manner with the preceding one and finely discriminant for large data bases. It is globally built by normalization with respect to an interesting set of association rules. Then the new index is contextual. The resulting increase of the computing complexity remains linear with respect to the size of the latter rule set.

This conceptual construction has been extremely extended in order to compare mutually in a symmetric way a set of complex attributes observed on a training object set. For this purpose, a given descriptive attribute is interpreted in terms of a discrete or weighted binary relation on the object set $\mathcal{O}$ (see [22] and the associated references). Qualitative attributes of any sort are included in this generalization. This takes an essential part in the development of the *LLA* hierarchical classification method, when the latter adresses the problem of the classification of the attribute set [20].

Otherwise, in the building process of the local likelihood link implicative measure, we have obtained in a coherent way a new absolute measure $\eta_3(a,b)$. This coherence consideration is also situated in the framework of the χ^2 theory. For a given data base observe that the contextual probabilistic indices $\mathcal{J}_n(a^j,a^k)$ [cf. (61)] can be obtained from the indices $\eta_3(a^i,a^k)$ by means of an increasing function.

However the specificity of the numerical values of $\mathcal{J}_n$ with respect to those of η_3 enables precisely to better distinguish between different association rules according to the unlikelihood principle. Note that this principle constitutes a basis of the information theory philosophy. Moreover, $\mathcal{J}_n$ enables easily to compare (relatively) a given association rule $a \to b$ in the contexts of two different sets of rules. Finally and mathematically, several indices conceived locally lead to the unique probabilistic index $\mathcal{J}_n$ by the global likelihood of the link construction (cf 4.1).

The experimental validation includes different interesting aspects. Our main goal was to compare the respective behaviours of the local and the global probabilistic indices. In this, we have clearly shown the discriminant ability of the global normalized index. For this purpose the classical "mushrooms" data base has been employed. It would be interesting next to try other data bases as *car, votes, monks* and other adata bases available on `http://www.ics.uci.edu/~mlearn/`.

More importantly, it is interesting in the future to continue the experimental analysis by studying with the same experimental scheme the behaviour of other indices such that those listed in [27], $\psi(a,b)$, called "entropic intensity" or $\eta_3(a,b)$, mentionned above.

Finally, when using $\mathcal{J}_n$ in order to evaluate the interest of a given rule in the context of a data base, boundaries could be defined by the expert knowledge. For this purpose one can for example consider inclusion situations such as the one provided in Figure 5.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami (1993) Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207–216
2. Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lofti Lakhal (2002) Pascal : un algorithme d'extraction des motifs fréquents. *Techniques et Science Informatiques*, 21(1):65–95
3. J. P. Bénzecri (1973) Théorie de l'information et classification d'après un tableau de contingence In *L'Analyse des Donnnées, Tome 1 : La Taxinomie*, Dunod, 207– 236
4. J-M. Bernard and C. Charron (1996) L'analyse implicative bayésienne, une méthode pour l'étude des dépendances orientées: Données binaires. *Revue Mathématique Informatique et Sciences Humaines*, 134:5–38
5. C.L. Blake and C.J. Merz (1998) UCI repository of machine learning databases
6. S. Brin, R. Motwani, and C. Silverstein (1997) Beyond market baskets: generalizing association rules to correlations. In *Proceedings of ACM SIGMOD'97*, 265–276
7. F. Daudé (1992) *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. PhD thesis, Université de Rennes 1
8. R. M. Goodman and P. Smyth (1998) Information-theoretic rule induction. In *ECAI 1988*, 357–362
9. R. Gras (1979) *Contribution à l'étude expérimentale et l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. PhD thesis, Doctorat d' État, Université de Rennes 1
10. R. Gras, P. Kuntz, and H. Briand (2001) Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données. *Revue Mathématique et Sciences Humaines*, 154-155:9–29
11. R. Gras and A. Larher (1992) L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématique Informatique et Sciences Humaines*, 18(120):5–31
12. S. Guillaume (2000) *Traitement des données volumineuses. Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. PhD thesis, Université de Nantes
13. P. Jaccard (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise en Sciences Naturelles*, 44:223–270

14. S. Lallich and O. Teytaud (2004) Évaluation et validation de l'intérêt des règles d'association. In *Mesures de Qualité pour la Fouille des Données*, RNTI-E-1, Cépaduès 193 – 218
15. R. Lehn (2000) *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. PhD thesis, Institut de Recherche en Informatique de Nantes
16. P. Lenca, P. Meyer, B. Vaillant, B. Picouet, and S. Lallich (2004) Évaluation et analyse multicritère des mesures de qualité des règles d'association. In *Mesures de Qualité pour la Fouille des Données*. Cépaduès 219 – 245
17. I. C. Lerman (1984) Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publications de l'Institut de Statistique des Universités de Paris*, 29:27–57
18. I. C. Lerman (1991) Foundations of the likelihood linkage analysis (lla) classification method. *Applied Stochastic Models and Data Analysis*, 7:63–76
19. I.C. Lerman (1970) Sur l'analyse des données préalable à une classification automatique ; proposition d'une nouvelle mesure de similarité. *Mathématiques et Sciences Humaines*, 8:5–15
20. I.C. Lerman (1981) *Classification et analyse ordinale des données*. Dunod
21. I.C. Lerman (1993) Likelihood linkage analysis (lla) classification method (around an example treated by hand). *Biochimie*, 75:379–397
22. I.C. Lerman (2006) Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. application des données génotypiques. *Revue de Statistique Appliquée*, in press
23. I.C. Lerman, R. Gras, and H. Rostam (1981) Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Revue Mathématique et Sciences Humaines*, 74 and 75:5–35, 5–47
24. J. Loevinger (1947) A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61:1–49
25. K. Pearson (1900) On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. In *Philosophical Magazine*, 157 – 175
26. G. Piatetsky-Shapiro (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, MIT Press 229 – 248
27. P-N. Tan, V. Kumar, and J. Srivastava (2002) Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*