Parametrised measures for the evaluation of association rules interestingness

Stéphane Lallich[1], Benoît Vaillant[2], and Philippe Lenca[2]

[1] Laboratoire ERIC, Université Lyon 2, France
(e-mail: stephane.lallich@univ-lyon2.fr)
[2] GET - ENST Bretagne -CNRS TAMCIC, France
(e-mail: forname.name@enst-bretagne.fr)

Abstract. In this paper, we present a original and synthetical overview of most commonly used association rule interestingness measures. These measure usually relate the confidence of a rule to an independency reference situation. Others relate it to indetermination, or impose a minimum confidence threshold. We propose a systematic generalisation of these measures, taking into account the reference point choosen by an expert in order to apprehend the confidence of a rule. This generalisation introduces new connections between measures, leads to the enhancement of some of them, and we propose new parametrised possibilities.
Keywords: interestingness measure, independency, indetermination.

1 Motivations

In this paper, we focus on the generalisating objective interestingness measures. We will consider association rule intersetingness measures, which aim at quantifying the quality of rules extracted from binary transactional datasets. In such datasets, each row is representing an object of the data mined, and consists of binary attributes, relating each object with properties that it may have or not. In this context, an association rule is an implication $\mathtt{A} \rightarrow \mathtt{B}$, where $\mathtt{A}$ and $\mathtt{B}$ (also called *itemsets*) are conjunctions of attributes. We denote by n the total number of transactions in the database, n_a (resp. n_b, n_{ab}, $n_{a\bar{b}}$) the number of transactions matching $\mathtt{A}$ (resp. $\mathtt{B}$, $\mathtt{A}$ and $\mathtt{B}$, $\mathtt{A}$ but not $\mathtt{B}$), and by p_a (resp. p_b, p_{ab}, $p_{a\bar{b}}$) the corresponding relative frequencies. Most objective measures are expressed as real valued functions of n, of the marginal frequencies p_a, p_b, and either p_{ab} or $p_{a\bar{b}}$, *i.e.* as functions of n, and of the *confidence* (Conf) p_{ab}/p_a and marginal frequency counts of the considered rule since $p_{a\bar{b}} = p_a - p_{ab}$. The higher the value of the measure, the better the rule is expected to be. Considering that the more counter-examples to a rule there are, the worst it is, we restrict our set of measures to those decreasing with $p_{a\bar{b}}$ (see table 1, references may be found in [Lenca *et al.*, 2004]). For a larger list of measures the reader should refer to [Guillet, 2004].

Support (Sup) and confidence (Conf) are the most famous of such measures, being the fundamentals principles of Apriori-like algorithms [Agrawal

and Srikant, 1994]. These algorithms extract rules such that their SUP and CONF is above given thresholds, σ_s and σ_c. They are deterministic [Freitas, 2000], and produce a large number of rules which may not be interessting:

- one would expect from a rule that its CONF should be above a reference value, but the later seldom if ever equals σ_c. Two main references are clearly identified as worthy from a user point of view. The first one is p_b, which corresponds to the independence of the itemsets `A` and `B`. In this case the user wishes to focus on rules such that the prior knowledge of `A` increases the knowledge of `B`, *i.e.* rules having a confidence $p_{b/a}$ above the *a priori* frequency p_b. An alternative reference sometimes used is 0.5, as in [Blanchard *et al.*, 2005]. In our opinion, the first reference is to be taken within a *targeting* strategy, and the second one when considering a *predictive* strategy. More generally, a user may be interested in taking into account a reference value θ, $0 < \sigma_c \leq \theta \leq 1$, and will consider only rules having a CONF greater than θ. Fukuda Gain (FUKU) is an example of such a measure, where $\theta = \sigma_c$.
- what is more, the data mined is often subject to some sampling scheme. In order to take that into account, a special kind of measures have been proposed. They are called "statistical" in the sense that, unlike the others (also called "descriptive" measures), their value rises with n, the relative frequencies being fixed. This consideration accounts for developping an inferential approach, and retaining only rules that are significantly well evaluated by measures, comparison to the reference choosen. Amongst the issues that arise from this approach, validating a large number of rules through the control of false rules discovery is assessed in [Lallich *et al.*, 2004].

Various properties of interestingness measures have been investigated, in particular in [Piatetsky-Shapiro, 1991], [Hilderman and Hamilton, 1999], [Freitas, 1999], [Lallich, 2002], [Lallich and Teytaud, 2004], [Gras *et al.*, 2004] and [Lenca *et al.*, 2004]. One of these properties deals with the reference value to which the measure compares confidence, that is to say p_b (independency), 0.5 (indetermination), or some other value.

In this paper, we present a general survey of association rule interestingness measures and parametrise the reference value to which the measures will compare the confidence of a rule in order to estimate its quality. Such a consideration leads to an organised review of classical measures, the introduction of new ones, and enables us to enhance the coherence of some of them. We will first focus on descriptive measures, and then look at the statistical ones.

2 Descriptives measures

2.1 Reference to independency

Amongst frequently used measures added to SUP and CONF in order to capture the interestingness of a rule, are those taking the independence of the

	Authors	Relative definitions
Sup	(Agrawal and Srikant, 1994)	p_{ab}
Conf	(Agrawal and Srikant, 1994)	$p_{b/a}$
r	(Pearson, 1896)	$\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_{\bar{a}} p_b p_{\bar{b}}}}$
CenConf		$p_{b/a} - p_b$
PS	(Piatetsky-Shapiro, 1991)	$np_a\left(p_{b/a} - p_b\right) = np_a p_b\,(\text{Lift} - 1)$
Loe	(Loevinger, 1947)	$\frac{p_{b/a}-p_b}{p_{\bar{b}}} = \frac{1}{p_{\bar{b}}}\text{CenConf} = 1 - \frac{1}{\text{Conv}}$
- ImpInd	(Lerman *et al.*, 1981)	$\sqrt{n}\frac{p_{a\bar{b}}-p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$
Lift	(Brin *et al.*, 1997)	$\frac{p_{b/a}}{p_b}$
LC	(Azé and Kodratoff, 2002)	$\frac{p_{ab}-p_{a\bar{b}}}{p_b} = 2\frac{p_a}{p_b}\,(\text{Conf} - 0.5)$
Seb	(Sebag and Schoenauer, 1988)	$\frac{p_{ab}}{p_{a\bar{b}}} = \frac{\text{Conf}}{1-\text{Conf}}$
OM	(Jeffreys, 1935)	$\frac{p_{b/a}/p_{\bar{b}/a}}{p_b/p_{\bar{b}}} = \frac{p_{ab}}{p_b}\frac{p_{\bar{b}}}{p_{a\bar{b}}} = \text{Lift}\cdot\text{Conv}$
Conv	(Brin *et al.*, 1997)	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$
ECR		$1 - p_{a\bar{b}}/p_{ab} = 1 - 1/Seb$
IG	(Church and Hanks, 1990)	$\log\frac{p_{ab}}{p_a p_b} = \log(\text{Lift})$
IntImp	(Gras *et al.*, 1996)	$P\left[Poi\left(np_a p_{\bar{b}}\right) \geq np_{a\bar{b}}\right]$
EII	(Gras *et al.*, 2001)	$\left\{\left[\left(1-h_1(p_{ab})^2\right)\left(1-h_2(p_{ab})^2\right)\right]^{1/4}\varphi\right\}^{1/2}$
PDI	(Lerman and Azé, 2003)	$P\left[\mathcal{N}(0,1) > \text{ImpInd}^{RC/\mathcal{B}}\right]$
Fuku	(Fukuda *et al.*, 1996)	$np_a\left(p_{b/a} - \sigma_c\right)$
Gan	(Ganascia, 1988)	$2p_{b/a} - 1$

- $h_1(t) = -(1-\frac{t}{p_a})\log_2(1-\frac{t}{p_a}) - \frac{t}{p_a}\log_2(\frac{t}{p_a})$ if $t \in [0, p_a/2[$; else $h_1(t) = 1$
- $h_2(t) = -(1-\frac{t}{p_{\bar{b}}})\log_2(1-\frac{t}{p_{\bar{b}}}) - \frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}})$ if $t \in [0, p_{\bar{b}}/2[$; else $h_2(t) = 1$
- Poi stands for Poisson and $\mathcal{N}(0,1)$ for the standard normal distribution
- ImpInd $^{CR/\mathcal{B}}$ corresponds to ImpInd, centred reduced (CR) for a rule set $\mathcal{B}$

Table 1. List of measures

itemsets `A` and `B` as reference. This is the case of many linear transformation of Conf: the centered confidence (CenConf), Piatetsky-Shapiro (PS), Loevinger (Loe), the implication index (ImpInd), and the lift (Lift). All these measures additively centre confidence on p_b from $p_{b/a} - p_b$, save Lift for which the centring is multiplicative and based on $\frac{p_{b/a}}{p_b}$. Other monotonically increasing transformations of confidence making reference to independency are the odd multiplier ($OM = \frac{1-p_b}{p_b} \times \frac{Conf}{1-Conf}$), the conviction ($Conv = \frac{1-p_b}{1-Conf}$), whereas the information gain ($IG = \log Lift$) is a transformation of Lift.

2.2 Reference to indetermination

Some measures may (explicitly or not) refer to the indetermination situation, when the number of examples and counter-examples is balanced for a given n_a [Blanchard *et al.*, 2005]. This is the case of Conf and the two linear transformation: least confidence ($LC = 2 \times (p_{b/a} - 0.5) \times \frac{p_a}{p_b}$) and the Ganascia measure ($Gan = 2 \times (Conf - 0.5)$) that both additively centre Conf at 0.5. Other transformations can be listed, in particular the Sebag and Shoenauer measure ($Seb = \frac{Conf}{1-Conf}$) and the examples and counter-examples rate ($ECR = \frac{2\times(Conf-0.5)}{Conf}$).

2.3 Reference at θ

In order to generalise the expression of interestingness measures with respect to θ, *i.e.* rules such that $1 \geq Conf(\mathtt{A} \rightarrow \mathtt{B}) \geq \theta(\mathtt{A} \rightarrow \mathtt{B})$, we will alternatively consider the quantities $Conf - \theta$, $\frac{Conf}{\theta}$ and $\frac{Conf-\theta}{1-\theta}$. Descriptive interestingness measures are generalised as follows:

$CenConf_{|\theta} = Conf - \theta$
$Gan_{|\theta} = \frac{Conf-\theta}{1-\theta} = Loe_{|\theta} = \frac{1}{1-\theta} CenConf_{|\theta}$
$Fuku_{|\theta} = PS_{|\theta} = np_a\,(Conf - \theta)$
$Lift_{|\theta} = \frac{Conf}{\theta}$
$IG_{|\theta} = \log(Lift_{|\theta})$
$Conv_{|\theta} = \frac{1-\theta}{1-Conf}$
$OM_{|\theta} = Seb_{|\theta} = \frac{Conf}{\theta} \times \frac{1-\theta}{1-Conf} = Lift_{|\theta} \times Conv_{|\theta}$
$LC_{|\theta} = \frac{Conf-\theta}{1-\theta} \times \frac{p_a}{p_b} = Loe_{|\theta} \times \frac{p_a}{p_b}$

Some measures in table 1 are particular instances of several generalised expression:

$$OM_{|\theta=p_b} = Seb_{|\theta=0.5}, \quad Gan_{|\theta=0.5} = Loe_{|\theta=p_b}, \quad Fuku_{|\theta=\sigma_c} = PS_{|\theta=p_b}$$

3 Statistical measures

3.1 Intrinsics of statistic and probabilistic measures

As mentioned previously, a statistic measure takes into account the size of the sampling scheme. It is qualified of "probabilistic" when expressed as the complement of the p-value of the test under $p_{b/a} \leq p_b$ hypothesis. Classical approaches use the independence of itemsets $\mathtt{A}$ and $\mathtt{B}$ hypothesis as reference. The modelling of this hypothesis realised in [Lerman *et al.*, 1981] can be done in three different ways, with respectively 1, 2 and 3 hazard levels. We introduce model $1'$ which is an alternative to model 1 where p_a is fixed, rather than n_a (table 2).

We denote by N_{ab} the random variable generating n_{ab}, and H and B refer respectively to the hypergeometric and binomial laws. The statistic and probabilistic index based on $n_{a\bar{b}}$ are built as follows: by establishing the law of N_{ab} et $N_{a\overline{b}}$ under null hypothesis (H_0) following the choosen modelling, we can express a centered and reduced index under H_0, noted $N^{CR}_{a\overline{b}}$. Under standard conditions, the law of this index can be approximated to the normal distribution, leading to the definition of a probabilistic measure, defined as the surprise of observing such a high value of the index under H_0. The choosen modelling does not affect the expectation, but does modify the variance. [Gras, 1979] and [Lerman *et al.*, 1981] prefer the third modelling, that dissociates most rules $\mathtt{A} \rightarrow \mathtt{B}$ and $\overline{\mathtt{B}} \rightarrow \overline{\mathtt{A}}$ whereas the first modelling makes no dinstinction between these rules. The measure hence obtained is the implication intensity (IntImp), which is most satisfying on properties one expects a measure should have [Lenca *et al.*, 2004], [Gras *et al.*, 2004].

	Modelling 1 and 1′	Modelling 2	Modelling 3
Principle	1.1 n_a fixed, N_{ab} randomised 1.1' p_a fixed N_{ab} randomised	2.1 $N_a \equiv B(n, p_a)$ 2.2 $/N_a = n_a$, $N_{ab} \equiv B(n_a, p_b)$	3.1 $N \equiv P(n)$ 3.2 $/N = n$, $N_a \equiv B(n, p_a)$ 3.3 $/N = n$, $N_a = n_a$, $N_{ab} \equiv B(n_a, p_b)$
Law N_{ab} under H_0	1.1 $H(n, n_a, p_b)$ 1.1' $B(n_a, p_b)$	$B(n, p_a p_b)$	$Poi(np_a p_b)$
Law $N_{a\overline{b}}$ under H_0	1.1 $H(n, n_a, p_{\overline{b}})$ 1.1' $B(n_a, p_{\overline{b}})$	$B(n, p_a p_{\overline{b}})$	$Poi(np_a p_{\overline{b}})$
Statistical index N_{ab}^{CR}	1.1 $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{a}} p_b p_{\overline{b}}}}$ $= -r\sqrt{n}$ 1.1' $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_b p_{\overline{b}}}}$	$\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{b}}(1 - p_a p_{\overline{b}})}}$	$IndImp$ $\frac{N_{a\overline{b}} - np_a p_{\overline{b}}}{\sqrt{np_a p_{\overline{b}}}}$
Probabilistic index $P(N(0,1) > N_{a\overline{b}}^{CR})$	1.1 $P(N(0,1) < r)$		IntImp $P(N(0,1) > IndImp)$

Table 2. Modelling of the various statistical and probabilistic index

3.2 Retaining the discriminating power

Although having many good properties, one of the major drawbacks of IntImp (drawback shared by the other statistic and probabilistic measures) is the loss of discriminating power. By its definition, it will evaluate rules significantly different from independency between 0.95 and 1. If n becomes important, which is particularly true in a data mining context, the slightest divergence from an independency situation becomes highly significant, thus leading to high and homogeneous values of the measure, close to 1.

In order to counter-balance this loss in discriminating power, [Lerman and Azé, 2003] introduce a contextual approach where ImpInd is centered and reduced (^{CR} notation) on a case database $\mathcal{B}$, thus leading to the definition of the probabilistic discriminant index, $PDI = P\left[N(0,1) > ImpInd^{CR/\mathcal{B}}\right]$.

[Gras *et al.*, 2001] propose an alternative solution by wheighting IntImp through the use of an inclusion index. This index is based on the entropy of experiments $\mathtt{B}/\mathtt{A}$ and $\overline{\mathtt{A}}/\overline{\mathtt{B}}$. We denote by $H(X) = p_x \log_2 p_x + p_{\overline{x}} \log_2 p_{\overline{x}}$ the entropy associated with an event X. In [Blanchard *et al.*, 2004] the most general form of the inclusion index is given as:

$i(A \subset B) = \left[(1 - H^*(B/A)^\alpha)\left(1 - H^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$

where $H^*(X) = H(X)$ if $p_x > 0.5$, $H^*(X) = 1$ otherwise. The α parameter is choosen by the user. The value $\alpha = 2$ is advised if one wants that this index should be tolerant to initial counter-examples, and we will use this value from now on. Hence, [Gras *et al.*, 2001] define the entropic intensity of implication as $EII = [IntImp \cdot i(A \subset B)]^{\frac{1}{2}}$

The shift from $H(X)$ to $H^*(X)$ aims at discarding uninteresting situations, such as $p_{b/a} < 0.5$ or $p_{\overline{a}/\overline{b}} < 0.5$, and complies with a predictive strategy. In a targeting strategy, the value of $p_{b/a}$ should have been compared to p_b, and the value of $p_{\overline{a}/\overline{b}}$ to $p_{\overline{a}}$.

The wheighting of the implication of intensity by the inclusion index, although effective, is problematic. The inclusion index is a measure of the distance to indetermination based on entropy, thus being null when $p_{b/a} = 0.5$, and so is EII. Still, IntImp values 0.5 at independency. Hence EII is not always null at independency: $EII = \sqrt[8]{\frac{(1-H(A)^2)(1-H(B)^2)}{16}}$ if $p_a < 0.5$ and $p_b > 0.5$, and is null otherwise.

3.3 Revised entropic intensity of implication

We will now propose two adaptations if EII in order to cope with the above mentioned issues: *REII* (Revised EII) et TEII (Troncated EII). Our first proposal consists in replacing IntImp by $IntImp^* = \max\{2IntImp - 1; 0\}$ in EII. This will solve the issues pointed out, but has the inconvenient of modifying the entire spectrum of values taken by EII:

$REII = [IntImp^* \cdot i(A \subset B)]^{\frac{1}{2}}$

Our second proposal solely nullifies the values of EII when $\frac{n_a n_{\overline{b}}}{n} \leq n_{a\overline{b}} \leq \min\{\frac{n_a}{2}, \frac{n_{\overline{b}}}{2}\}$, whithout modifying its values otherwise. To achieve this, we introduce an adequate version of $H(X)$. In order to take into account both predictive and targeting strategies, a rule will have a null evaluation by the inclusion index, and hence by TEII when the following conditions are jointly met:

- $p_{b/a} > 0.5$ (prediction) and $p_{b/a} > p_b$ (targeting); *i.e.* $p_{b/a} > \max(0.5, p_b)$
- $p_{\overline{a}/\overline{b}} > 0.5$ (prediction) and $p_{\overline{a}/\overline{b}} > p_{\overline{a}}$ (targeting); *i.e.* $p_{\overline{a}/\overline{b}} > \max(0.5, p_{\overline{a}})$

With these new conditions, TEII is null whenever the number of counter-examples is above $\min\left(\frac{n_a n_{\overline{b}}}{n}; \frac{n_a}{2}; \frac{n_{\overline{b}}}{2}\right)$. $TEII = [IntImp(A \to B) \times i_t(A \subset B)]^{\frac{1}{2}}$, with:

- $i_t(A \subset B) = \left[(1 - H_t^*(B/A)^\alpha)\left(1 - H_t^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$,
- $H_t^*(B/A) = H(B/A)$ if $p_{b/a} > \max(0.5, p_b)$, $H_t^*(B/A) = 1$ otherwise,
- $H_t^*(\overline{A}/\overline{B}) = H(\overline{A}/\overline{B})$ if $p_{\overline{a}/\overline{b}} > \max(0.5, p_{\overline{a}})$, $H_t^*(\overline{A}/\overline{B}) = 1$ otherwise.

3.4 Measures making reference to indetermination

[Blanchard *et al.*, 2005] propose IPEE, a probabilistic measure of deviation from equilibrium. The authors implicitly use modelling $1'$ since they consider $N_{a\overline{b}} \equiv B(n_a, 0.5)$ under indetermination hypothesis, *i.e.* $N_{a\overline{b}}^{CR} = \frac{N_{a\overline{b}} - 0.5n_a}{0.5\sqrt{n_a}}$. They introduce $IPEE = P\left[B(n_a, 0.5) > n_{a\overline{b}}\right] \approx P\left[N(0,1) > \frac{n_{a\overline{b}} - 0.5n_a}{0.5\sqrt{n_a}}\right]$. Under normal approximation, IPEE equals 0.5 at indetermination. This measure corresponds to the probalistic index associated to modelling $1'$ (see table 2), where p_b is replaced by 0.5. IPEE will hence inherit of the weak discriminating power of this kind of measures, thus leading the authors to propose that it should be modulated by the inclusion index, which is all the most coherent, since both index make reference to indetermination.

3.5 Generalised intensity of implication

Using the same approach as with descriptive measures, we can generalise statistical measures and evaluate the interestingness of a rule by comparing its CONF to θ. This is done by considering in table 2 that for each modelling under H_0, the probability of an example, conditionally to n_a, of an example is θ: $N_{ab} \equiv B(n_a, \theta)$.

The results of the hence adapted modelling 1 is immediate, and those of modelling 2 and 3 are easily obtained through the use of the probability generating functions. If $X \equiv B(m,p)$, its generating function then is $G(s) = E(s^X) = (1-p+ps)^m$, and if $X \equiv Poi(\lambda)$, it is $G(s) = E(s^X) = e^{-\lambda(1-s)}$.

- In modelling 2, n is fixed, $N_a \equiv B(n, p_a)$ and $N_{ab}/(N_a = n_a) \equiv B(n_a, \theta)$. Since $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E\left((1-\theta+\theta s)^{N_a}\right)$, we have:
$$N_{ab} \equiv B(n, \theta p_a) \text{ and } N_{a\bar{b}} \equiv B(n, (1-\theta)p_a)$$
- In modelling 3, we have $N \equiv Poi(n)$, $N_a/(N=n) \equiv B(n,p_a)$, and $N_a/(N=n \text{ and } N_a = n_a) \equiv B(n_a, \theta)$.
As $G_{N_a}(s) = E(s^{N_a}) = E(E(s^{N_a}/N)) = E((1-p_a+p_as)^N) = e^{-np_a(1-s)}$, then $N_a \equiv Poi(np_a)$.
Similarly, since $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E((1-\theta+\theta s)^{N_a}) = e^{-n\theta p_a(1-s)}$, we have:
$$N_{ab} \equiv Poi(n\theta p_a) \text{ and } N_{a\bar{b}} \equiv Poi(n(1-\theta)p_a)$$

From these results, we propose a range of generalised measures (see table 1), and will focus on two of them. The first one, $GIPE_{|\theta}$, associated to modelling $1'$ and generalises IPEE. It corresponds to the χ^2 adjustment of B/A distribution and $(\theta;\ 1-\theta)$. The second one, $GIntImp_{|\theta}$, associated to modelling 3 generalises INTIMP.

	Modelling 1 and 1′	Modelling 2	Modelling 3
Principle	1.1 n_a fixed, N_{ab} randomised 1.1' p_a fixed, N_{ab} randomised	2.1 $N_a \equiv B(n,p_a)$ 2.2 $/N_a = n_a$, $N_{ab} \equiv B(n_a,\theta)$	3.1 $N \equiv Poi(n)$ 3.2/$N=n$, $N_a \equiv B(n,p_a)$ 3.3 $/N=n$, $N_a = n_a$, $N_{ab} \equiv B(n_a,\theta)$
Law N_{ab}	1.1 $H(n,n_a,\theta)$ 1.1' $B(n_a,\theta)$	$B(n,\theta p_a)$	$Poi(np_a\theta)$
Law $N_{a\bar{b}}$	1.1 $H(n,n_a,1-\theta)$ 1.1' $B(n_a,1-\theta)$	$B(n,(1-\theta)p_a)$	$Poi(np_a(1-\theta))$
Statistical index $N^{CR}_{a\bar{b}}$	1.1 $\frac{N_{a\bar{b}}-np_a(1-\theta)}{\sqrt{np_ap_{\overline{a}}\theta(1-\theta)}}$ 1.1' $\frac{N_{a\bar{b}}-np_a(1-\theta)}{\sqrt{np_a\theta(1-\theta)}}$	$\frac{N_{a\bar{b}}-np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}$	$GIndImp_{\|\theta}$ $\frac{N_{a\bar{b}}-np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$
Probabilistic index $P(N(0,1) > N^{CR}_{a\bar{b}})$	1.1' $GIPE_{\|\theta}$		$GIntImp_{\|\theta} = P(N(0,1) > GIndImp_{\|\theta})$

Table 3. Modelling of the various generalised index

3.6 Discriminant power of the generalised measures

The generalised statistical or probabilistic measures have, as the original ones do, a weak discriminating power. In order to enhance these measures, we will consider two approaches, one being contextual, like [Lerman and Azé, 2003], the other one relying on a weighting through the use of an inclusion index, like [Gras *et al.*, 2001].

In the contextual approach, GINDIMP $_{|\theta}$ (or its equivalent following modelling 1 and 2) is centred and reduced on a case database $\mathcal{B}$, and thus define a generalised probabilistic discriminant index, GIPD $_{|\theta}$, as follows.

$$GIPD_{|\theta} = P\left(N(0,1) > GIndImp_{|\theta}^{CR/\mathcal{B}}\right)$$

This way, we also define the generalised entropic intensity of implication, $\text{GEII}_{|\theta}$, as the product of $\text{GINDIMP}_{|\theta}$ and an inclusion index. In order to remain coherent, we think advisable to use a generalised inclusion index $i_{|\theta}$, using θ as reference value and not 0.5. This can be achieved by replacing in the original formula $H(B/A)$ by $\widetilde{H}_{|\theta}(B/A)$ and $H(\overline{A}/\overline{B})$ by $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ where:

- $\widetilde{H}_{|\theta}(B/A)$ is expressed as $H(B/A)$, in which we replace $p_{b/a}$ by $\widetilde{p}_{b/a}$ defined as follows:

$$\widetilde{p}_{b/a} = \frac{p_{b/a}}{2\theta} \text{ if } p_{b/a} \leq \theta,\ \widetilde{p}_{b/a} = \frac{p_{b/a} + 1 - 2\theta}{2(1-\theta)} \text{ otherwise}$$

- $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ can be expressed either:
  - by considering θ as reference, in which case we form $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ as we did for $\widetilde{H}_{|\theta}(B/A)$, by replacing $p_{\overline{a}/\overline{b}}$ by $\widetilde{p}_{\overline{a}/\overline{b}}$ in $H(\overline{A}/\overline{B})$, with:

$$\widetilde{p}_{\overline{a}/\overline{b}} = \frac{p_{\overline{a}/\overline{b}}}{2\theta} \text{ if } p_{\overline{a}/\overline{b}} \leq \theta,\ \widetilde{p}_{\overline{a}/\overline{b}} = \frac{p_{\overline{a}/\overline{b}} + 1 - 2\theta}{2(1-\theta)} \text{ otherwise}$$

    This first possibility generalises the inclusion index proposed in [Gras *et al.*, 2001], and can be found back using $\theta = 0.5$.
  - or using $1 - \frac{p_a}{p_{\overline{b}}} \times (1-\theta)$ as reference, since $p_{\overline{a}/\overline{b}} = 1 - \frac{p_a}{p_{\overline{b}}} \times (1 - p_{b/a})$. In this case, when considering independancy (*i.e.* $\theta = p_b$), the reference value for $\widetilde{H}_{|\theta}(\overline{A}/\overline{B})$ is $p_{\overline{a}}$.

$\widetilde{H}^*_{|\theta}(B/A)$ and $\widetilde{H}^*_{|\theta}(\overline{A}/\overline{B})$, are defined as:

$$\widetilde{H}^*_{|\theta}(X) = \widetilde{H}_{|\theta}(X) \text{ if } p_x > \theta,\ \widetilde{H}^*_{|\theta}(X) = 1 \text{ otherwise}$$

and $i_{|\theta}$ as:

$$i_{|\theta} = \left[\left(1 - \widetilde{H}^*_{|\theta}(B/A)^\alpha\right)\left(1 - \widetilde{H}^*_{|\theta}(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}, \text{ with } \alpha = 2.$$

From this, we deduce $\text{GEII}_{|\theta}$ as $GEII_{|\theta} = \left[IntImp_{|\theta} \times i_{|\theta}\right]^{\frac{1}{2}}$, which is a more discriminating version of GIntImp. A similar approach leads to the definition of a generalised probabilistic measure of deviation, $\text{GEIPE}_{|\theta}$, as $GEIPE_{|\theta} = \left[GIPE_{|\theta} \times i_{|\theta}\right]^{\frac{1}{2}}$.

Their behaviour, compared to their original counterparts, is represented figure 1. They were obtained using 3 different values for θ, $\theta = 0.1$ (thus targeting at independency), $\theta = 0.2$ (targeting for situations such that `B` happens twice more often when `A` is true) and $\theta = 0.5$ (prediction).

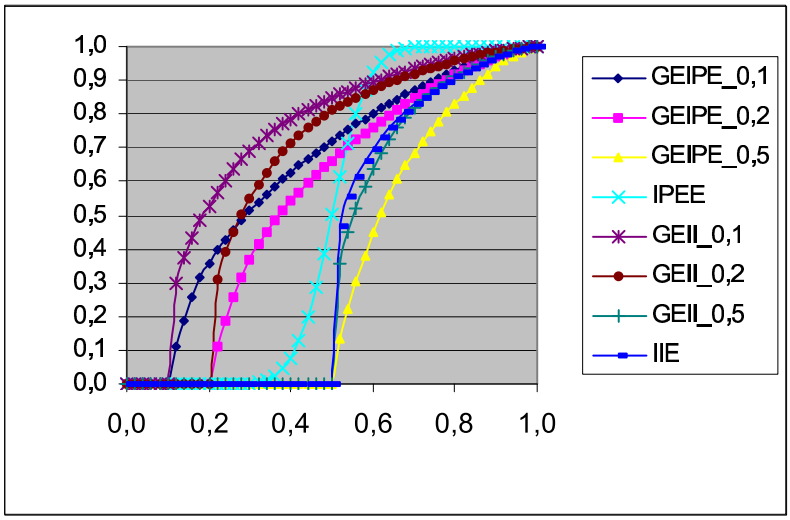

Fig. 1. Behaviour of the measures, in function of $p_{b/a}$ for $n = 1000$, $p_a = 0.05$ and $p_b = 0.10$

4 Conclusion

Following modelling and coherence principles, we proposed in this paper an innovating framework, from which a unified view of a large number of interestingness measures can be drawn, and which clarifies some of the links between these measures. Moreover, this framework is at the basis of the definition of new measures, namely the generalised intensity of implication, generalised probabilistic discriminant index, generalised entropic intensity of implication and the generalised probabilistic measure of deviation from equilibrium, that all compare the confidence of a rule to a reference parameter.

References

[Agrawal and Srikant, 1994]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[Blanchard *et al.*, 2004]J. Blanchard, P. Kuntz, , F. Guillet, and R. Gras. Mesure de la qualité des règles d'association par l'intensité d'implication entropique. *Revue des Nouvelles Technologies de l'Information*, 1(RNTI-E):33–43, 2004.

[Blanchard *et al.*, 2005]J. Blanchard, F. Guillet, H. Briand, and R. Gras. IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. In *Atelier DKQ*, pages 26–34, 2005.

[Freitas, 1999]A. Freitas. On rule interestingness measures. *Knowledge-Based Systems journal*, pages 309–315, 1999.

[Freitas, 2000]A. Freitas. Understanding the crucial differences between classification and discovery of association rules - a position paper. In *ACM SIGKDD Explorations*, volume 2, pages 65–69, 2000.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage (EGC 2001)*, 1(1-2):69–80, 2001.

[Gras *et al.*, 2004]R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. Quelques critères pour une mesure de qualité de règles d'association. *RNTI-E-1*, 2004.

[Gras, 1979]R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* PhD thesis, Université de Rennes I, 1979.

[Guillet, 2004]F. Guillet. Mesure de la qualité des connaissances en ECD. In *Tutoriel de la 4e Conf. Extraction et Gestion des Connaissances (60 p.)*, 2004.

[Hilderman and Hamilton, 1999]Robert J. Hilderman and Howard J. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report 99-4, Dpt. of Computer Science, University of Regina, october 1999.

[Lallich and Teytaud, 2004]S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des règles d'association. *RNTI-E*, 1:193–217, 2004.

[Lallich *et al.*, 2004]S. Lallich, E. Prudhomme, and O. Teytaud. Contrôle du risque multiple en sélection de règles d'association significatives. In *EGC 04*, volume 2, pages 305–316, 2004.

[Lallich, 2002]S. Lallich. Mesure et validation en extraction des connaissances à partir des données. HDR – Université Lyon 2, 2002.

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. *RNTI-E*, 1:219–246, 2004.

[Lerman and Azé, 2003]I.C. Lerman and J. Azé. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *RSTI-RIA (EGC 2003)*, 1(17):247–262, 2003.

[Lerman *et al.*, 1981]I.C. Lerman, R. Gras, and H. Rostam. Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75):5–35, 5–47, 1981.

[Piatetsky-Shapiro, 1991]G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.