Using information-theoretic measures to assess association rule interestingness

Julien Blanchard, Fabrice Guillet, Régis Gras, Henri Briand

Using Information-theoretic Measures to Assess Association Rule Interestingness

Julien Blanchard, Fabrice Guillet, Régis Gras, Henri Briand
LINA (FRE 2729 CNRS) – Polytechnic School of Nantes University
La Chantrerie BP 50609 – 44306 Nantes cedex 3 – France
{julien.blanchard,fabrice.guillet,henri.briand,regis.gras}@polytech.univ-nantes.fr

Abstract

Assessing rules with interestingness measures is the cornerstone of successful applications of association rule discovery. However, there exists no information-theoretic measure which is adapted to the semantics of association rules. In this article, we present the Directed Information Ratio *(DIR), a new rule interestingness measure which is based on information theory. DIR is specially designed for association rules, and in particular it differentiates two opposite rules $a \rightarrow b$ and $a \rightarrow \bar{b}$. Moreover, to our knowledge, DIR is the only rule interestingness measure which rejects both independence and (what we call) equilibrium, i.e. it discards both the rules whose antecedent and consequent are negatively correlated, and the rules which have more counter-examples than examples. Experimental studies show that DIR is a very filtering measure, which is useful for association rule post-processing.*

1 Introduction

Many data mining techniques produce results in the form of rules. These are expressions of the type "if *antecedent* then *consequent*" where the boolean propositions *antecedent* and *consequent* are conjunctions of assignment expressions *variable=value*. Rules have the advantage of being very intelligible for users since they model information explicitly. They are also a major element of most theories of knowledge representations in cognitive sciences [10], and in particular they underlie many works in artificial intelligence, such as the expert systems. In knowledge discovery in databases, the main rule-based paradigms are the classification rules, used in supervised learning to predict a unique class variable as consequent, and the association rules [1], which can have any combination of variables as antecedent and consequent. Classification rules can be generated by induction algorithms such as CN2 [9] or decision tree algorithms such as C4.5 [18], while association rules are mined by combinatorial algorithms such has *Apriori* [1].

Due to their unsupervised nature, association rule mining algorithms commonly generate large amounts of rules, with much redundancy [25]. To help the user to find relevant knowledge in this mass of information, one of the main solutions consists in evaluating and sorting the rules with interestingness measures. There are two kinds of measures: the subjective (user-oriented) ones and the objective (data-oriented) ones. Subjective measures take into account the user's goals and domain knowledge [14] [16], whereas only the data cardinalities appear in the calculation of objective measures (surveys can be found in [22], [11], [24], [2]). In this article, we are interested in the objective measures. We have shown in [4] that there are two different, but complementary, aspects of the rule interestingness: the deviation from independence and the deviation from what we call *equilibrium* (maximum uncertainty of the consequent given that the antecedent is true). Thus, the objective measures of interestingness can be classified into two classes:

- the measures of deviation from independence, which have a fixed value when the antecedent and consequent are independent ($p(ab) = p(a).p(b)$), such as lift [2], conviction [8], rule-interest [17], Loevinger index [15], implication intensity [6];
- the measures of deviation from equilibrium, which have a fixed value when examples and counter-examples are equal in numbers ($p(ab) = p(a\bar{b}) = \frac{1}{2}p(a)$), such as confidence [1], Sebag and Schoenauer index [19], IPEE [4].

Among the objective measures of rule interestingness, the information-theoretic measures are particularly intelligible and useful since they can be interpreted in terms of information. More precisely, as pointed out by Smyth and Goodman [21], there is an interesting parallel to draw between the use of information theory [20] in communication systems and the use of information theory to evaluate rules. In communication systems, a channel has a high capacity if it can carry a great deal of information from the source to the receiver. As for a rule, the relation is interesting when the antecedent provides a great deal of information about

conditional entropy (H_c)	$-p_{b/a}.\log_2 p_{b/a} - p_{\overline{b}/a}.\log_2 p_{\overline{b}/a}$
mutual information (MI)	$p_{ab}.\log_2 \frac{p_{ab}}{p_a p_b} + p_{a\overline{b}}.\log_2 \frac{p_{a\overline{b}}}{p_a p_{\overline{b}}}$ $+ p_{\overline{a}b}.\log_2 \frac{p_{\overline{a}b}}{p_{\overline{a}} p_b} + p_{\overline{a}\overline{b}}.\log_2 \frac{p_{\overline{a}\overline{b}}}{p_{\overline{a}} p_{\overline{b}}}$
Theil uncertainty coefficient (u)	$\frac{MI}{-p_b.\log_2 p_b - p_{\overline{b}}.\log_2 p_{\overline{b}}}$
J-measure (J)	$p_{ab}.\log_2 \frac{p_{ab}}{p_a p_b} + p_{a\overline{b}}.\log_2 \frac{p_{a\overline{b}}}{p_a p_{\overline{b}}}$
Gini index (G)	$p_a({p_{b/a}}^2 + {p_{\overline{b}/a}}^2) +$ $p_{\overline{a}}({p_{b/\overline{a}}}^2 + {p_{\overline{b}/\overline{a}}}^2) - {p_b}^2 - {p_{\overline{b}}}^2$

Table 1. Information-theoretic measures of interestingness for a rule $a \rightarrow b$

the consequent (Smyth and Goodman speak of the information content of a rule [21]). The information-theoretic measures commonly used to evaluate rule interestingness are the Shannon conditional entropy [9], the average mutual information [12] (often simply called mutual information), the Theil uncertainty coefficient [23] [22], the J-measure [21], and the Gini index [2] [12] (cf. the formulas in table 1). The Shannon conditional entropy measures the average amount of information of the consequent given that the antecedent is true (it is used in the CN2 algorithm). The average mutual information (Shannon entropy decrease) measures the average information shared by the antecedent and the consequent. The Theil uncertainty coefficient measures the entropy decrease rate of the consequent due to the antecedent. The J-measure is the part of the average mutual information relative to the truth of the antecedent. Finally, the Gini index is the quadratic entropy decrease.

Even if these measures are commonly used to evaluate association rules (see [11], [22], [2]), they are all better suited to evaluate classification rules. As pointed out by Jaroszewicz and Simovici [12], an association rule should be assessed only on the variable values which are comprised in the rule[1], whereas the information-theoretic measures consider the full joint distribution of the antecedent and consequent (this is relevant for classification rules since in supervised learning, the user is interested in all the values of the consequent because it is the class variable). Consequently, the information-theoretic measures do not vary with the permutation of the values of a variable[2]. This invariance is undesirable for association rules since the permutation of values definitely transforms an association rule.

[1]Indeed, association rule mining algorithms transform each multivalued variable into several binary variables called "items".

[2]More precisely, the Shannon conditional entropy and the J-measure vary with the permutation of the values of a variable in the antecedent, but not in the consequent.

We say that association rules are not only "variable-based" relations but also "value-based" relations. If all the same such measures are applied on association rules, then this must be done carefully since it is not possible to distinguish between positive and negative correlations [22].

To be appropriate to association rules, an interestingness measure must respect their value-based semantics by not systematically giving the same value to a rule $a \rightarrow b$ and to its opposite $a \rightarrow \overline{b}$. Intuitively, if $a \rightarrow b$ is strong, then $a \rightarrow \overline{b}$ should be weak. In this article, we propose an interestingness measure based on information theory which respects the value-based semantics of association rules. This new measure named *DIR* (for *Directed Information Ratio*) allows to reject both the independence and equilibrium situations, i.e. with only one fixed threshold it allows to discard both the rules whose antecedent and consequent are negatively correlated, and the rules which have more counter-examples than examples. To our knowledge, this is a unique feature for a rule interestingness measure. In the next section, we introduce the new measure DIR from earlier works on the assessment of rules using information theory. Section 3 will then review the properties of DIR. Finally, in section 4 we compare DIR to other rule interestingness measures within the framework of formal and experimental studies.

2 Measuring the information content of rules

2.1 Notations

We consider a set of objects described by boolean variables. In the association rule terminology, the objects are transactions stored in a database, the variables are called items, and the conjunctions of variables are called itemsets. An association rule is a couple (a, b) noted $a \rightarrow b$ where a and b are two itemsets which have no items in common. The examples of the rule are the objects which verify the antecedent a and the consequent b, while the counter-examples are the objects which verify a but not b. A rule is all the better since it has lots of examples and few counter-examples. In the following, we study two itemsets a and b that we simply call the variables.

The Shannon entropy of the variable a is:

$$H(a) = -p(a=1).\log_2 p(a=1) - p(a=0).\log_2 p(a=0)$$

The Shannon conditional entropy of the variable b given an event $a = 1$ is defined by:

$$\begin{aligned} H(b/a=1) \quad = \quad & -p(b=1/a=1).\log_2 p(b=1/a=1) \\ & -p(b=0/a=1).\log_2 p(b=0/a=1) \end{aligned}$$

As can be seen, the entropic functions combine variables and realizations of variables. In order to distinguish them,

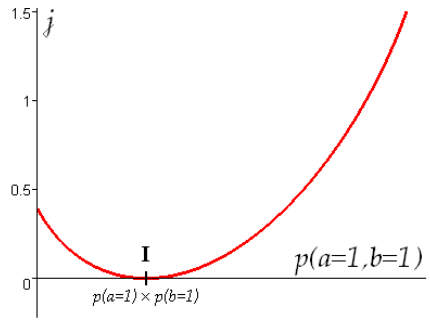


Figure 1. Plot of the measure j w.r.t. $p(a=1,b=1)$

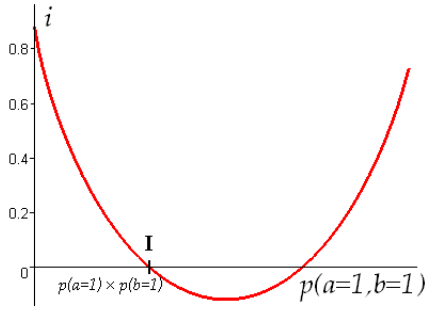


Figure 2. Plot of the measure i w.r.t. $p(a=1,b=1)$

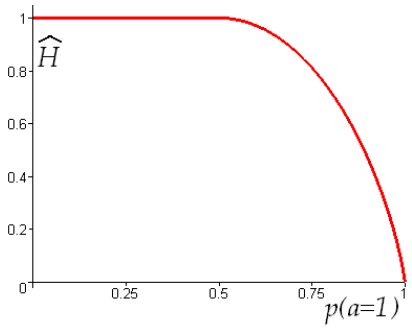


Figure 3. Plot of the reduced entropy $\widehat{H}(a)$

(I refers to the statistical independence of a and b)

the realizations of a variable b must be noted $b=1$ and $b=0$ in this article, and not b and $\overline{b}$ as commonly done in the association rule literature. With these explicit notations, an association rule should be written $(a=1) \rightarrow (b=1)$, but we retain the classical notation $a \rightarrow b$.

2.2 The amount of information that $a=\alpha$ gives about b

Let us consider the amount of information that an event $a=\alpha$ gives about a variable b ($\alpha \in \{0;1\}$). We note $M(a=\alpha,\ b)$ the measures of this amount of information. Blachman [3] studied the $M(a=\alpha,\ b)$ whose expectation (when averaged over all α) is the average mutual information between the variables a and b:

$$MI(a,b) = \mathrm{E}_\alpha\{M(a=\alpha,\ b)\} \qquad (1)$$

The two most frequently used measures are the following (see figures 1 and 2):

$$\begin{aligned} j(a=\alpha,\ b) &= p(b=1/a=\alpha).\log_2 \tfrac{p(b=1/a=\alpha)}{p(b=1)} \\ &+ p(b=0/a=\alpha).\log_2 \tfrac{p(b=0/a=\alpha)}{p(b=0)} \\ i(a=\alpha,\ b) &= H(b) - H(b/a=\alpha) \end{aligned}$$

Blachman shows that j is the only non-negative information-theoretic measure satisfying (1), while i is the only antisymmetric[3] information-theoretic measure satisfying (1).

The measure j is the cross-entropy between the *a priori* and *a posteriori* distributions of b. It is traditionally accepted as "the" measure of the amount of information that $a=\alpha$ gives about b. In particular, the J-measure (the most commonly used information-theoretic measure within the context of association rules) directly comes from j: $J = j \times \mathrm{P}(a=\alpha)$ [21]. Although the meaning of the measure i is much more obvious (it is the entropy decrease of b due to the event $a=\alpha$), one prefers j to i because j vanishes only if the variables a and b are independent, while i can vanish outside the independence (see figures 1 and 2). This behavior is due to the symmetrical nature of the entropy H (it does not vary with the permutation of the variable values).

2.3 Reduced entropy

In order to remove the symmetry introduced by the entropy in the measure i, we propose to use a directed entropic function $\widehat{H}$ named *reduced entropy* [5] (see figure 3).

Definition 1 The **reduced entropy** $\widehat{H}(a)$ of a variable a is defined by:

- if $p(a=1) \leq \frac{1}{2}$ then $\widehat{H}(a)=1$,
- if $p(a=1) \geq \frac{1}{2}$ then $\widehat{H}(a)=H(a)$.

One similarly defines the conditional reduced entropy of the variable b given the realization of a:

- if $p(b=1/a=1) \leq \frac{1}{2}$ then $\widehat{H}(b/a=1)=1$,
- if $p(b=1/a=1) \geq \frac{1}{2}$ then $\widehat{H}(b/a=1)=H(b/a=1)$.

The entropy $H(a)$ of a variable a can be written as the sum of two reduced entropies:

$$H(a) = \widehat{H}(a) + \widehat{H}(\overline{a}) - 1, \text{ with } \overline{a} \text{ being the negation of } a.$$

[3] i is antisymmetric with regard to the *a priori* and *a posteriori* distributions $P = \{p(b)\}$ and $Q = \{p(b/a=\alpha)\}$ of the variable b: $i(P,Q) = -i(Q,P)$

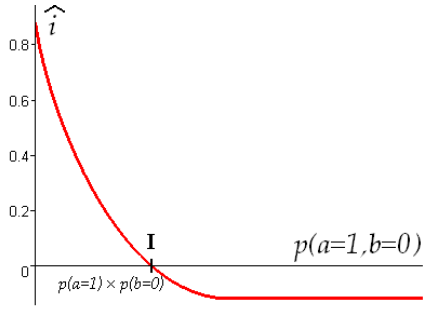


Figure 4. Plot of $\widehat{i}$ w.r.t. $p(a = 1, b = 0)$
(I refers to the statistical independence of a and b)

Contrary to H, $\widehat{H}$ is an asymmetric measure which differently evaluates an imbalance in favor of $a = 1$ and an imbalance in favor of $a = 0$: $\widehat{H}(a) \neq \widehat{H}(\overline{a})$. More precisely, if $a = 1$ is more frequent than $a = 0$, then:

- the reduced entropy $\widehat{H}(a)$ measures the entropy of a: $\widehat{H}(a) = H(a)$;
- the reduced entropy $\widehat{H}(\overline{a})$ is 1.

If $a = 1$ is less frequent than $a = 0$, then the roles are reversed. In other words, $\widehat{H}$ measures a "directed uncertainty" in favor of one of the values, in the sense that if this value is not the more likely, then the uncertainty is considered as maximal.

2.4 Directed Information Ratio

By introducing the reduced entropy $\widehat{H}$ in the measure i, we have:

$$i(a = 1,\ b) = \widehat{H}(b) + \widehat{H}(\overline{b}) - \widehat{H}(b/a = 1) - \widehat{H}(\overline{b}/a = 1)$$

Hence:

$$i(a = 1,\ b) = \widehat{i}(a = 1,\ b)\ +\ \widehat{i}(a = 1,\ \overline{b})$$

$$\text{with } \widehat{i}(a = 1,\ b) = \widehat{H}(b) - \widehat{H}(b/a = 1)$$

So the index i which measures the decrease of the entropy H is the sum of two decreases of reduced entropy $\widehat{H}$:

- $\widehat{i}(a = 1,\ b)$ which is the decrease of reduced entropy of b due to $a = 1$,
- $\widehat{i}(a = 1,\ \overline{b})$ which is the decrease of reduced entropy of $\overline{b}$ due to $a = 1$.

Contrary to the measures i and j, the new index $\widehat{i}(a = 1,\ b)$ is absolutely appropriate to evaluate the interestingness of an association rule $a \rightarrow b$:

$$\widehat{i}(a = 1,\ b) = \widehat{i}(a \rightarrow b)$$

Indeed, $\widehat{i}(a = 1,\ b)$ increases with the number of examples (probability $p(a = 1, b = 1)$), decreases with the number of counter-examples (probability $p(a = 1, b = 0)$, see figure 4), and respects the value-based semantics of association rules by differentiating opposite rules $a \rightarrow b$ and $a \rightarrow \overline{b}$. The higher $\widehat{i}(a = 1,\ b)$, the more the event $a = 1$ brings information in favor of $b = 1$, and the more the interestingness of the rule $a \rightarrow b$ is guaranteed. If $\widehat{i}(a = 1,\ b)$ is negative, this means that $a = 1$ brings no information in favor of $b = 1$, and even that it "removes" some information (the uncertainty is lesser by predicting $b = 1$ randomly rather than by predicting $b = 1$ using the rule $a \rightarrow b$). In our opinion, $\widehat{i}$ is a measure of what Smyth and Goodman call the information content of rules [21].

Like the directed contribution to χ^2 [13], $\widehat{i}$ allows to distribute the average mutual information of two variables over the rules between them:

$$\begin{aligned} MI(a,b) &= p(a=1).\widehat{i}(a \rightarrow b) + p(a=1).\widehat{i}(a \rightarrow \overline{b}) \\ &+ p(a=0).\widehat{i}(\overline{a} \rightarrow b) + p(a=0).\widehat{i}(\overline{a} \rightarrow \overline{b}) \end{aligned}$$

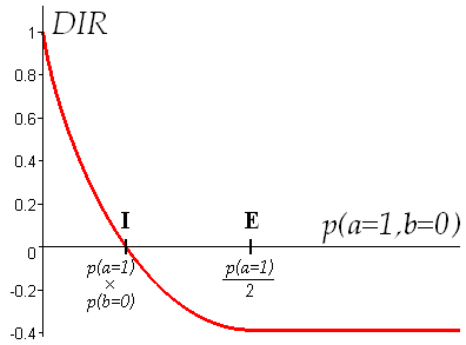
$p(a = 1).\widehat{i}(a \rightarrow b)$ is the directed contribution of the rule $a \rightarrow b$ to the average mutual information. Each rule takes part in the average mutual information by giving or removing its share of infomation. Like the χ^2, the average mutual information can also be written with the contributions of the four opposite rules.

For all these characteristics, we propose to retain the index $\widehat{i}$ to evaluate the interestingness of association rules. However, a drawback of $\widehat{i}$ is that its maximal value is not fixed but depends on $p(b = 1)$, making the comparison of rules with different consequents difficult. This maximal value is obtained for logical rules, i.e. rules with no counter-examples ($p(a = 1, b = 0) = 0$). In order to facilitate the filtering of the most informative rules, we normalize $\widehat{i}$ by assigning the maximal value 1 to the logical rules. This amounts to calculating the decrease rate of reduced entropy.

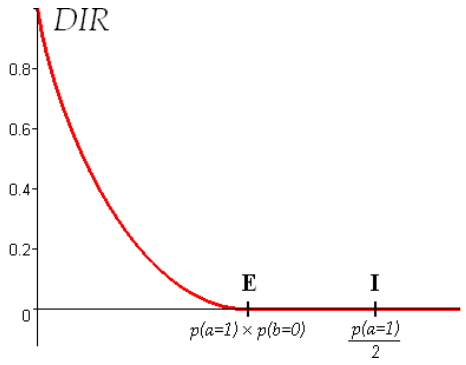
Definition 2 The **Directed Information Ratio** (DIR) of a rule $a \rightarrow b$ is defined by:

$$DIR(a \rightarrow b) = \frac{\widehat{H}(b) - \widehat{H}(b/a = 1)}{\widehat{H}(b)} \quad \text{if } p(b = 1) \neq 1$$

If $p(b = 1) = 1$, then $\widehat{H}(b) = 0$ and DIR is not defined. However, such rules are obviously to be discarded since they are completely expected ($\widehat{i}$ is indeed 0 for these rules). A rule is said to be informative if its DIR is strictly positive.



(A) $p(b=1) \geq \frac{1}{2}$



(B) $p(b=1) \leq \frac{1}{2}$

Figure 5. Plot of *DIR* w.r.t. $p(a=1,b=0)$
(I refers to independence and E to equilibrium)

3 *DIR* properties

The main properties of DIR are given in table 2. It must be noticed that DIR satisfies the three properties that define a good interestingness measure according to Piatetsky-Shapiro [17]: it is 0 at independence, it increases with the examples, and it decreases with the sizes of the antecedent and consequent (variations with all other parameters fixed). Furthermore, DIR has no symmetry:

- it does not assign the same value to $a \rightarrow b$ and to its opposite $a \rightarrow \overline{b}$, since it respects the value-based semantics of association rules;
- it does not either assign the same value to $a \rightarrow b$ and to its converse $b \rightarrow a$, which is better when the user interprets association rules as quasi-implications [22].

As shown in figure 5, DIR is a convex decreasing function of the number of counter-examples. Among the rule interestingness measures, it belongs to the demanding indexes, i.e. the indexes which decrease quickly with the first counter-examples and thus allow to better discriminate the good rules (larger dispersion of values).

Range	$]-\infty\ ;\ 1]$
Value for logical rules	1
Value for independence	0
Value for equilibrium	$1-\widehat{H}(b)^{-1} \leq 0$
Variation w.r.t. $p(a=1,b=1)$	$\nearrow$
Variation w.r.t. $p(a=1)$	$\searrow$
Variation w.r.t. $p(b=1)$	$\searrow$

Table 2. *DIR* properties

Let us consider a rule $(a \rightarrow b)$ described by the probabilities $p(a=1)$, $p(b=1)$, and $p(a=1,b=0)$ [4]. The independence is defined by $p(a=1,b=0) = p(a=1).p(b=0)$, while the equilibrium is defined by $p(a=1,b=0) = \frac{1}{2}p(a=1)$. By varying $p(a=1,b=0)$ with fixed $p(a=1)$ and $p(b=1)$, one can distinguish two different cases for DIR:

- If $p(b=1) \geq \frac{1}{2}$, then
  $p(a=1).p(b=0) \leq \frac{1}{2}p(a=1)$
  so the rule goes through the independence before going through the equilibrium when $p(a=1,b=0)$ increases. The measure DIR vanishes at independence and then admits negative values (figure 5.(A)).

- If $p(b=1) \leq \frac{1}{2}$, then
  $p(a=1).p(b=0) \geq \frac{1}{2}p(a=1)$
  so the rule goes through the equilibrium before going through the independence when $p(a=1,b=0)$ increases. The measure DIR vanishes but does not admit negative values (figure 5.(B)).

DIR allows to reject both the independence and equilibrium situations. Indeed, in these situations, DIR is either negative or worth zero (see table 2). By retaining only strictly positive values of DIR (informative rules), the user discards all the rules whose deviation from indepedence is bad (rules between negatively correlated variables), and also all the rules whose deviation from equilibrium is bad (rules with more counter-examples than examples). So, the measure must be used with a strictly positive threshold to filter the rules. To our knowledge, DIR is the only rule interestingness measure which can reject both independence and equilibrium with a fixed threshold. This approach is completely original for rule interestingness assessment.

[4] As often in the association rule literature, we choose the probability of counter-examples as a parameter, but the results are the same with the probability of examples since $p(a=1,b=1) = p(a=1) - p(a=1,b=0)$.

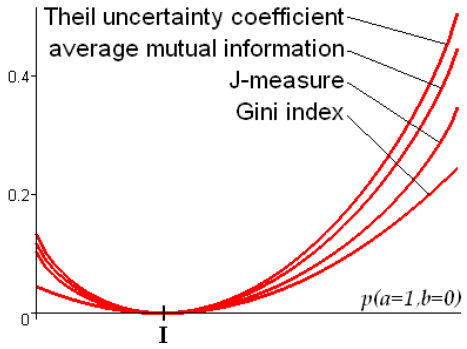


Figure 6. Information-theoretic measures of deviation from independence

4 Comparison to other measures

4.1 Formal comparison

In this section, we compare DIR to the information-theoretic measures traditionally used to evaluate rule interestingness (see table 1 for formulas):

- the Shannon conditional entropy [9], which measures the deviation from equilibrium;
- the mutual information [12], the Theil uncertainty [23] [22], the J-measure [21], and the Gini index [2] [12], which measure the deviation from independence.

As the last four measures have similar behaviors (see figure 6), we only plot one of them in the comparisons below. We choose the J-measure since it is used a lot within the context of association rules (in particular it does not assign the same value to a rule $a \rightarrow b$ and to its converse $b \rightarrow a$). As for the conditional entropy, it is not the function H_c of the table 1 which is plotted in the comparisons below, but the complementary function $1 - H_c$. Indeed, H_c assigns its smallest values to the best rules[5]. One generally prefers the opposite behavior for a rule interestingness measure [17].

The figures 7.(A) and 7.(B) compare DIR to the conditional entropy and to the J-measure when the probability of counter-examples $p(a = 1, b = 0)$ increases. The figures clearly illustrate that the conditional entropy and the J-measure do not respect the value-based semantics of association rules, since they can increase even though the probability of counter-examples increases. Moreover, one can see that DIR and the conditional entropy have the advantage of systematically assigning the value 1 to the logical rules, which are the best rules from an objective point of view. This makes the comparisons among the rules easier, and facilitates the choice of a threshold to filter the rules. On the contrary, for the J-measure and the three other measures of deviation from independence, a value can be assigned to a good rule (lots of examples, few counter-examples), even though on other data the same value would be assigned to a bad rule. In fact, except for the value 0 which always correponds to independence, the values taken by these measures cannot be interpreted in an absolute way, i.e. independently of the data.

The figures 7.(A) and 7.(B) show that the conditional entropy detects the equilibrium but not the independence (it could even take high values at independence). On the other hand, the J-measure detects the independence but not the equilibrium. In all cases, filtering the rules on DIR with a strictly positive threshold is enough to reject both independence and equilibrium. As illustrated in figure 7.(B), DIR is similar to the conditional entropy when $p(b = 1) \leq \frac{1}{2}$ (the functions are partly identical). This is what enables DIR to detect the equilibrium when $p(b = 1) \leq \frac{1}{2}$.

[5] To generate relevant rules, the CN2 algorithm tries to minimize H_c, and not to maximize it [9].

	Items	Objects	Outputted rules
T10.I4.D5k	12	5000	97688
T10.I4.D100k	1000	100000	478894
BREAKDOWNS	92	2883	43930
PROFILES	30	2299	28938

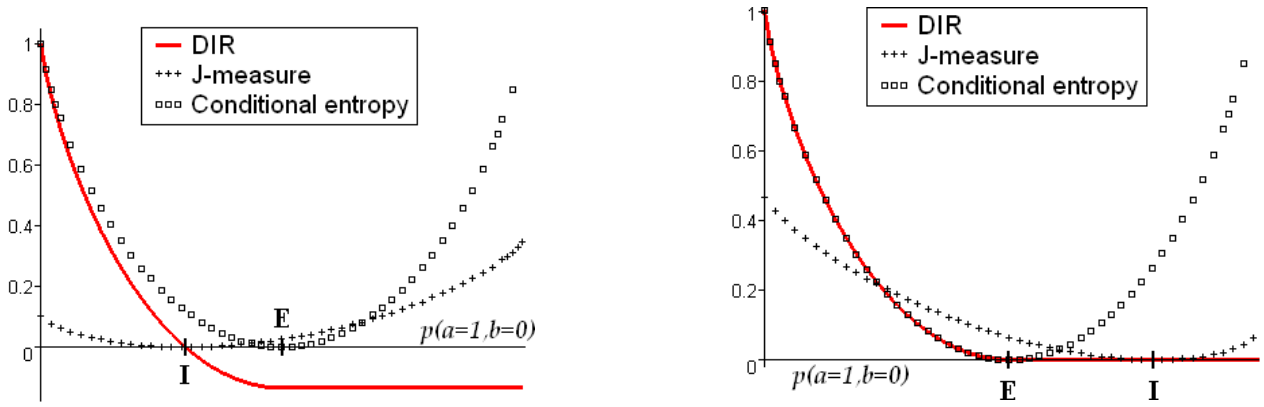
Table 3. Data characteristics

4.2 Experimental comparison

We compare the distributions of DIR to the distributions of other interestingness measures on the association rules mined from four datasets (described in table 3). The two first datasets were generated using the IBM synthetic data generator[6] described in [1] which simulates purchases in a supermarket. The two other datasets are a database of lift breakdowns provided by a lift maintenance company, and a database of workers' psychological profiles used in human resource management. The rules were mined with the *Apriori* algorithm [1] with a low support threshold to avoid the premature elimination of potentially interesting rules.

As we want here to compare the distributions of measures, we choose measures which, as DIR, take the value 1 for the logical rules. Among the information-theoretic measures, only the conditional entropy satisfies this condition. So, we add to our comparisons two reference measures of rule interestingness which satisfy the condition: the confidence [1] $(p(b = 1/a = 1))$ and the Loevinger index [15] $(1 - \frac{p(a=1,b=0)}{p(a=1).p(a=0)})$. They respectively measure the deviation from equilibrium and from independence. As figures

[6] http://www.almaden.ibm.com/software/quest/Resources/index.shtml



(A) $p(b=1) \geq \frac{1}{2}$ (B) $p(b=1) \leq \frac{1}{2}$

Figure 7. Plot of *DIR*, J-measure, and conditional entropy w.r.t. $p(a=1, b=0)$

8.(A-D) show, DIR is the most filtering index: for the four datasets, whichever the threshold chosen between 0 and 1, DIR prunes more rules than the others. This is especially useful within the context of association rules where the mining algorithms often generate huge amounts of rules.

Let us explain why DIR is very filtering. In figure 8.(E) in parallel coordinates, each line represents a rule. The figure shows representative rules from T10.I4.D5k that are judged good by confidence but not by the Loevinger index (they have a good deviation from equilibrium but not from independence). On the other hand, figure 8.(F) shows representative rules from BREAKDOWNS that are judged good by the Loevinger index but not by confidence (they have a good deviation from independence but not from equilibrium). DIR gives bad values to all these rules, since it takes into account both independence and equilibrium.

5 Conclusion

In this article, we have presented the *Directed Information Ratio* (DIR), a new rule interestingness measure which is based on information theory. DIR is specially designed for association rules, and in particular it respects their value-based semantics by differentiating the opposite rules $a \rightarrow b$ and $a \rightarrow \overline{b}$. Moreover, to our knowledge, DIR is the only rule interestingness measure which rejects both independence and equilibrium, i.e. it discards both the rules whose antecedent and consequent are negatively correlated, and the rules which have more counter-examples than examples. Experimental studies have also shown that DIR is a very filtering measure, which is useful for association rule post-processing. To continue this research work, we will integrate DIR into a data mining platform in order to experiment with this new measure in real applications.

Like all the information-theoretic measures, DIR is a frequential index. This means that it takes into account the size of the data only in an relative way, and not in an absolute way (see [4]). More generally, in order to have a complete assessment of the rules, one has to measure not only the deviations from equilibrium and independence, but also the statistical significance of these two deviations. For example, χ^2 [7] or implication intensity [6] allow to measure the statistical significance of the deviation from independence, while IPEE [4] allows to measure the statistical significance of the deviation from equilibrium. These approaches are complementary to DIR.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. pages 307–328. AAAI, 1996.

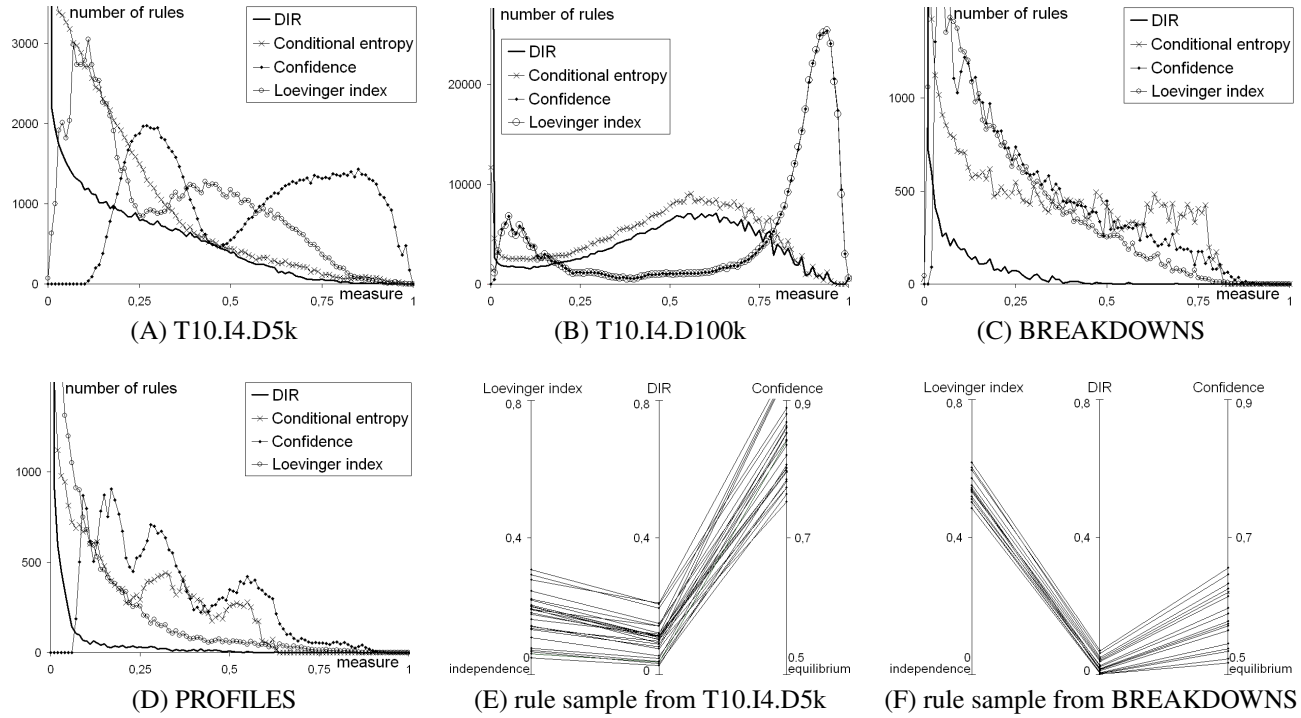
[2] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of ACM KDD'1999*, pages 145–154. ACM Press, 1999.

[3] N. M. Blachman. The amount of information that *y* gives about *x*. *IEEE Transcations on Information Theory*, IT-14(1):27–31, 1968.

[4] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, pages 191–200, 2005.

[5] J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. *Revue des Nouvelles Technologies de l'Information*, E-2:287–298, 2004. Actes EGC2004.

[6] J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. Implication intensity: from the basic statistical definition to the entropic version. In *Statistical Data Mining and Knowledge Discovery*, pages 473–485. Chapman & Hall, 2003. Chapter 28.



(A) T10.I4.D5k (B) T10.I4.D100k (C) BREAKDOWNS

(D) PROFILES (E) rule sample from T10.I4.D5k (F) rule sample from BREAKDOWNS

Figure 8. Measure distributions on the whole sets of rules (A, B, C, D) and on two samples of rules (E and F, in parallel coordinates)

[7] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. *SIGMOD Record*, 26(2):265–276, 1997.

[8] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record*, 26(2):255–264, 1997.

[9] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

[10] J. Holland, K. Holyoak, R. Nisbett, and P. Thagard. *Induction : Processes of inference, learning and discovery*. MIT Press, 1986.

[11] X.-H. Huynh, F. Guillet, and H. Briand. ARQAT: An exploratory analysis tool for interestingness measures. In *Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, pages 334–344, 2005.

[12] S. Jaroszewicz and D. A. Simovici. A general measure of rule interestingness. In *Proceedings of PKDD'2001*, pages 253–265. Springer-Verlag, 2001.

[13] I. Lerman. Foundations in the likelihood linkage analysis classification method. *Applied Stochastic Models and Data Analysis*, 7:69–76, 1991.

[14] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.

[15] J. Loevinger. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61(4), 1947.

[16] B. Padmanabhan and A. Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999.

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

[18] J. Quinlan, editor. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[19] M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proceedings of EKAW88*, pages 28.1–28.20, 1988.

[20] C. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.

[21] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, 1992.

[22] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[23] H. Theil. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76:103–154, 1970.

[24] B. Vaillant, P. Lenca, and S. Lallich. A clustering of interestingness measures. In *Proceedings of the 7th International Conference on Discovery Science*, pages 290–297, 2004.

[25] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9(3):223–248, 2004.