Assessing rule interestingness with a probabilistic measure of deviation from equilibrium

Julien Blanchard, Fabrice Guillet, Henri Briand, and Régis Gras

LINA – FRE 2729 CNRS
Polytech'Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Abstract. Assessing rule interestingness is the cornerstone of successful applications of association rule discovery. In this article, we present a new measure of interestingness named $IPEE$. It has the unique feature of combining the two following characteristics: first, it is based on a probabilistic model, and secondly, it measures the deviation from what we call *equilibrium* (maximum uncertainty of the consequent given that the antecedent is true). We study the properties of this new index and show in which cases it is more useful than a measure of deviation from independence.
Keywords: Data mining, Association rules, Interestingness measures, Statistical significance, Deviation from equilibrium.

1 Introduction

Among the knowledge models used in Knowledge Discovery in Databases (KDD), association rules [Agrawal *et al.*, 1993] have become a major concept and have received significant research attention. Association rules are implicative tendencies $a \rightarrow b$ where a and b are conjunctions of items (boolean variables of the form $databaseAttribute = value$). Such a rule means that if a record verifies the antecedent a in the database then it certainly verifies the consequent b.

A crucial step in association rule discovery is post-processing, i.e. the interpretation, evaluation, and validation of the rules in order to find interesting knowledge for decision-making. Indeed, due to their unsupervised nature, the data mining algorithms can produce a great many rules, many of which have no interest. To help the user (a decision-maker specialized in the data studied) 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 user's *a priori* knowledge of the data [Liu *et al.*, 2000] [Padmanabhan and Tuzhilin, 1999] [Silberschatz and Tuzhilin, 1996]. On

the other hand, only the data cardinalities appear in the calculation of objective measures [Tan *et al.*, 2004] [Bayardo and Agrawal, 1999] [Guillet, 2004] [Lenca *et al.*, 2004] [Lallich and Teytaud, 2004]. In this article, we are interested in the objective measures.

We have shown in [Blanchard *et al.*, 2004] that there exist 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 are divided into two groups:

- the measures of deviation from independence, which have a fixed value when the variables a and b are independent ($n.n_{ab} = n_a n_b$)[1];
- the measures of deviation from equilibrium, which have a fixed value when examples and counter-examples are equal in numbers ($n_{ab} = n_{a\bar{b}} = \frac{1}{2} n_a$).

The objective measures can also be classified according to their descriptive or statistical nature [Lallich and Teytaud, 2004] [Gras *et al.*, 2004]:

- The descriptive (or frequential) measures are those which do not vary with the cardinality expansion (when all the data cardinalities are increased or decreased in equal proportion).
- The statistical measures are those which vary with the cardinality expansion. Among them, one can find the probabilistic measures, which compare the observed distribution to an expected distribution, such as the implication intensity [Gras, 1996] [Blanchard *et al.*, 2003b] or the likelihood linkage index [Lerman, 1991].

	Measures of deviation from equilibrium	**Measures of deviation from independence**
Descriptive measures	– confidence, – Sebag et Schoenauer index, – example and counter-example ratio, – Ganascia index, – *moindre-contradiction*, – inclusion index...	– correlation coefficient, – lift, – Loevinger index, – conviction, – J-measure, – *TIC*, – odds ratio, – *multiplicateur de cote*...
Statistical measures		– implication intensity, – implication index, – likelihood linkage index, – oriented contribution to χ^2, – rule-interest...

Table 1. Classification of the objective measures of rule interestingness

With these two criteria, we classify the objective measures of rule interestingness into four categories. As shown in table 1 (cf. [Guillet, 2004] for

[1] The notations are defined in section 2

the references), there are no statistical measures which evaluate the deviation from equilibrium. Nevertheless, the statistical measures have the advantage of taking into account the size of the phenomena studied. Indeed a rule is statistically all the more reliable since it is assessed on a large amount of data. Moreover, when based on a probabilistic model, a statistical measure refers to an intelligible scale of values (a scale of probabilities); this is not the case for many interestingness measures. Also, such a measure facilitates the choice of a threshold for filtering the rules, since the complement to 1 of the threshold has the meaning of the significance level of a hypothesis test (generally in a test, one chooses $\alpha \in \{0.1\%, 1\%, 5\%\}$).

In this article, we propose a new measure of rule interestingness which evaluates the deviation from equilibrium while having a statistical nature. More precisely, this index is based on a probabilistic model and measures the statistical significance of the deviation from equilibrium (whereas implication intensity or likelihood linkage index, for example, measure the statistical significance of the deviation from independence). In the next section, we present a probabilistic index of deviation from equilibrium named *IPEE*, and then study in section 3 its properties. Section 4 is devoted to the comparison between the measures of deviation from equilibrium and the measures of deviation from independence.

2 Measuring the statistical significance of the deviation from equilibrium

We consider a set O of n 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. Given an itemset a, we note A the set of the objects which verify a, and n_a the cardinality of A. The complement of A in O is the set $\overline{A}$ of cardinality $n_{\overline{a}}$. 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 rule examples are the objects which verify the antecedent a and the consequent b (objects in $A \cap B$), while the rule counter-examples are the objects which verify a but not b (objects in $A \cap \overline{B}$). In the following, we call "variables" the itemsets.

2.1 Random model

Given a rule $a \rightarrow b$, we want to measure the statistical significance of the rule deviation from equilibrium. As the equilibrium configuration is defined by the equidistribution in A of examples $A \cap B$ and counter-examples $A \cap \overline{B}$, the null hypothesis is the hypothesis H_0 of equiprobability between the examples and counter-examples. So, let us associate to the set A a random set X of cardinality n_a drawn in O under this hypothesis: $\mathrm{P}(X \cap B) = \mathrm{P}(X \cap \overline{B})$ (cf. figure 1). The number of counter-examples expected under H_0 is the

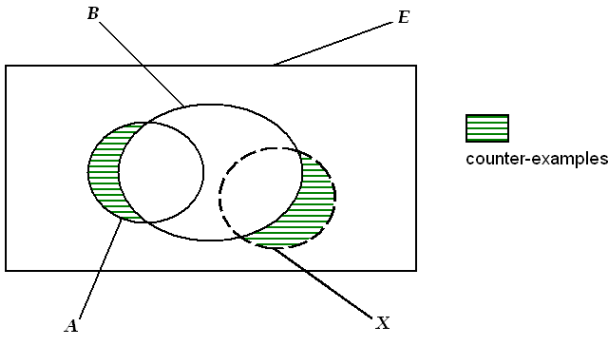


Fig. 1. Random draw of a set X under the equiprobability hypothesis between the examples and counter-examples

cardinality of $X \cap \overline{B}$, noted $\left|X \cap \overline{B}\right|$. It is a random variable whose $n_{a\bar{b}}$ is an observed value. The rule $a \rightarrow b$ is even better since there is a high probability that chance creates more counter-examples than data.

Définition 1 The **probabilistic index of deviation from equilibrium (*IPEE*[2])** of a rule $a \rightarrow b$ is defined by:

$$IPEE(a \rightarrow b) = \mathrm{P}(\left|X \cap \overline{B}\right| > n_{a\bar{b}} \ \mid \mathrm{H}_0)$$

A rule $a \rightarrow b$ is said to be acceptable with the confidence level $1 - \alpha$ if $\delta(a \rightarrow b) \geq 1 - \alpha$.

Therefore, *IPEE* quantifies the unlikelihood of the smallness of the number of counter-examples $n_{a\bar{b}}$ with respect to the hypothesis H_0. In particular, if $\delta(a \rightarrow b)$ is close to 1 then it is unlikely that the features (a *and* b) and (a *and* $\bar{b}$) are equiprobable. This new index can be seen as the complement to 1 of the p-value of a hypothesis test (and α as the significance level of this test). However, following the implication intensity and the likelihood linkage index (where H_0 is the hypothesis of independence between a and b), the aim is not testing a hypothesis but actually using it as a reference to evaluate and sort the rules.

2.2 Analytical expression

In the case of drawing random sets with replacement, $\left|X \cap \overline{B}\right|$ is binomial with parameters n_a and $\frac{1}{2}$:

$$\delta(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} \binom{n_a}{k}$$

[2] *IPEE* is for *Indice Probabiliste d'Ecart à l'Equilibre* in French

IPEE depends neither on n_b (it does not increase with the rarity of the consequent), nor on n since the equilibrium hypothesis H_0 is not defined by means of n_b and n (contrary to the independence hypothesis). It must be noticed that the statistical significance of the deviation from equilibrium could be measured by comparing not the counter-examples but the examples: $\widehat{IPEE}(a \rightarrow b) = P(|X \cap B| < n_{ab} \mid H_0)$. However, since the binomial distributions with parameter $\frac{1}{2}$ are symmetrical, the two indexes are identical:

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \binom{n_a}{n_a - K} = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \binom{n_a}{K} = \widehat{IPEE}(a \rightarrow b)$$

where $K = n_a - k$.

When $n_a \geq 10$, the binomial distribution can be approximated by the normal distribution with mean $\frac{n_a}{2}$ and standard deviation $\sqrt{\frac{n_a}{4}}$. The standardized number of counter-examples $\widetilde{n}_{a\bar{b}}$ can be interpreted as the oriented contribution to the χ^2 of goodness-of-fit between the observed distribution of examples/counter-examples, and the uniform distribution: $\chi^2 = \widetilde{n}_{a\bar{b}}{}^2$ This constitutes a strong analogy with the implication intensity and the likelihood linkage index, since in the poissonian models of these two measures, the standardized values of $n_{a\bar{b}}$ and n_{ab} can be seen as oriented contributions to the χ^2 of independence between a and b [Lerman, 1991].

3 *IPEE* properties

Range	$[0; 1]$
Value for logical rules	$1 - \frac{1}{2^{n_a}}$
Value for equilibrium	0.5
Variation w.r.t. $n_{a\bar{b}}$ with fixed n_a	$\searrow$
Variation w.r.t. n_a with fixed $n_{a\bar{b}}$	$\nearrow$

Table 2. *IPEE* properties

The properties and the graph of *IPEE* are given respectively in table 2 and figure 2. We can observe that :

- *IPEE* varies slightly with the first counter-examples (slow decrease). This behavior is intuitively satisfactory since a small number of counter-examples do not question a rule [Gras *et al.*, 2004].
- The discarding of the rules gets quicker in an uncertainty range around the equilibrium $n_{a\bar{b}} = \frac{n_a}{2}$ (fast decrease).

As shown in figure 3, with a ratio examples/counter-examples which is constant, the values of *IPEE* are all the more extreme (close to 0 or 1) since n_a is large. Indeed, owing to its statistical nature, the measure takes into account the size of the phenomena studied: the larger n_a is, the more one can trust the imbalance between examples and counter-examples observed in the data, and the more one can confirm the good or bad quality of the rule deviation from equilibrium. In particular, for *IPEE*, the quality of a logical rule (rule with no counter-examples, i.e. $n_{a\bar{b}} = 0$) depends on n_a (cf. table 2). Thus, contrary to the other measures of deviation from equilibrium (cf. table 1), *IPEE* has the advantage of systematically attributing the same value to the logical rules. This allows to differentiate and sort the logical rules.

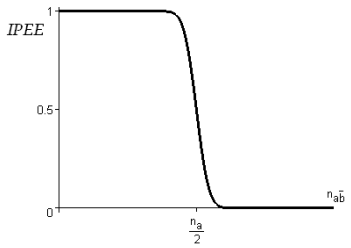


Fig. 2. Plot of *IPEE* w.r.t. the number of counter-examples $n_{a\bar{b}}$

It must be noticed that *IPEE* has no symmetry: it does not assign the same value to a rule $a \rightarrow b$ and to its converse $b \rightarrow a$, or to its contrapositive $\overline{b} \rightarrow \overline{a}$, or to its opposite $a \rightarrow \overline{b}$. Nevertheless, we have the following relation: $\delta(a \rightarrow \overline{b}) = 1 - \delta(a \rightarrow b) - \frac{C_{n_a}^{n_{ab}}}{2^{n_a}}$ (the last term is negligible when n_a is large).

We have seen that the strength of statistical significance measures lies in the fact that they take into account the size of the phenomena studied. On the other hand, it is also their main limit: the measures have a low discriminating power when the size of the phenomena is large (beyond around 10^4) [Elder and Pregibon, 1996]. Indeed, with regard to large cardinalities, even minor deviations can be statistically significant. *IPEE* does not depart from this: when n_a is large, the measure tends to evaluate the rules as either very good (values close to 1), or very bad (values close to 0). In this case, to fine-tune the filtering of the best rules, it is necessary to use a descriptive measure (cf. table 1) such as the inclusion index [Blanchard *et al.*, 2003b] in addition to *IPEE*. On the other hand, contrary to the implication intensity or the likelihood linkage index, *IPEE* does not depend on n. Therefore, the measure is sensitive to both the specific rules ("nuggets") and the general rules ; it can be used on either small or large databases.

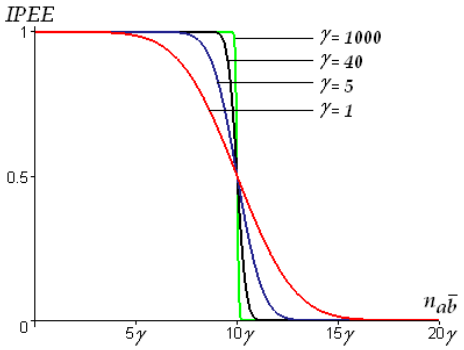


Fig. 3. Plot of *IPEE* w.r.t. cardinality expansion
($n_a = 20 \times \gamma,\ n_{a\bar{b}} \in [0 \times \gamma\ ;\ 20 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

4 Measures of deviation from equilibrium and independence: a comparison

Let us consider a rule with the cardinalities $n_{a\bar{b}}$, n_a, n_b, n. By varying $n_{a\bar{b}}$ with fixed n_a, n_b, and n, one can distinguish two different cases [Blanchard *et al.*, 2004] :

- If $n_b \geq \frac{n}{2}$ (case 1), then $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$, so the rule goes through the independence before going through the equilibrium when $n_{a\bar{b}}$ increases.
- If $n_b \leq \frac{n}{2}$ (case 2), then $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$, so the rule goes through the equilibrium before going through the independence when $n_{a\bar{b}}$ increases.

Let us now compare a measure of deviation from equilibrium M_{eql} and a measure of deviation from independence M_{idp} for these two cases. In order to have a fair comparison, we suppose that the two measures have similar behaviors:

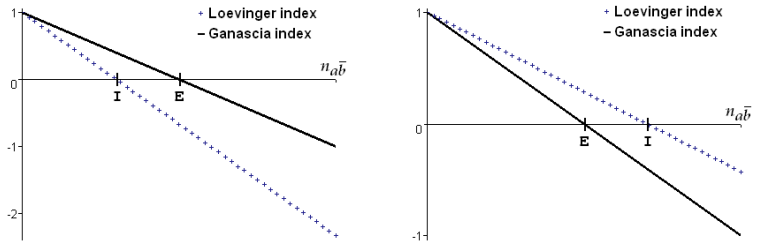
- same value for a logical rule,
- same value for equilibrium/independence,
- same decrease speed with regard to the counter-examples.

For example, M_{eql} and M_{idp} can be the Ganascia and Loevinger indexes [Ganascia, 1991] [Loevinger, 1947] (cf. the definitions in table 3), or *IPEE* and the implication intensity. As shown in figures 4 and 5, M_{idp} is more filtering than M_{eql} in case 1, whereas M_{eql} is more filtering than M_{idp} in case 2. In other words, in case 1, it is M_{idp} which contributes to rejecting the bad rules, while in case 2 it is M_{eql}. This confirms that the measures of deviation from equilibrium and the measures of deviation from independence have to be regarded as complementary, the second ones not being systematically "better" than the first ones. In particular, the measures of deviation from

equilibrium must not be neglected when the realizations of the studied variables are rare. Indeed, in this situation, should the user not take an interest in the rules having non-realizations (which is confirmed in practice), case 2 is more frequent than case 1.

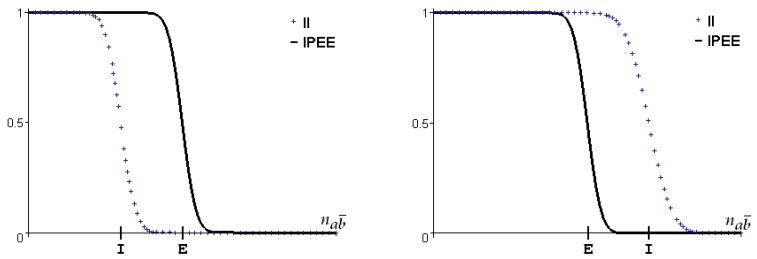
Ganascia index	Loevinger index
$\frac{2n_{ab}-n_a}{n_a}$	$1-\frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$

Table 3. Ganascia and Loevinger indexes for a rule $a \rightarrow b$



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 4. Comparison of the Ganascia and Loevinger indexes (E: equilibrium, I: independence)



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 5. Comparison of the measures *IPEE* and implication intensity (*II*)

5 Conclusion

In this article, we have presented a new measure of rule interestingness which evaluates the deviation from equilibrium with respect to a probabilistic model. Due to its statistical nature, this measure has the advantage of taking into account the size of the phenomena studied, contrary to the other measures of deviation from equilibrium. Moreover, it refers to an intelligible scale of values (a scale of probabilities). Our study shows that *IPEE* is efficient to assess logical rules, and well adapted to the search for specific rules ("nuggets").

IPEE can be seen as the counterpart of the implication intensity [Gras, 1996] [Blanchard *et al.*, 2003b] for the deviation from equilibrium. Used together, these two measures allow an exhaustive statistical evaluation of the rules. To continue this research work, we will integrate *IPEE* into our rule validation system *ARVis* [Blanchard *et al.*, 2003a] in order to experiment with the couple (*IPEE*, implication intensity) on real data.

References

[Agrawal *et al.*, 1993]Rakesh Agrawal, Tomasz Imielienski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, pages 207–216. ACM Press, 1993.

[Bayardo and Agrawal, 1999]Roberto J. Bayardo and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 145–154. ACM Press, 1999.

[Blanchard *et al.*, 2003a]Julien Blanchard, Fabrice Guillet, and Henri Briand. A user-driven and quality-oriented visualization for mining association rules. In *Proceedings of the third IEEE international conference on data mining ICDM'03*, pages 493–496. IEEE Computer Society, 2003.

[Blanchard *et al.*, 2003b]Julien Blanchard, Pascale Kuntz, Fabrice Guillet, and Régis Gras. Implication intensity: from the basic statistical definition to the entropic version. In Hamparsum Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 473–485. Chapman and Hall/CRC Press, 2003. chapter 28.

[Blanchard *et al.*, 2004]Julien Blanchard, Fabrice Guillet, Régis Gras, and Henri 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 des journées Extraction et Gestion des Connaissances (EGC) 2004.

[Elder and Pregibon, 1996]John F. Elder and Daryl Pregibon. A statistical perspective on knowledge discovery in databases. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 83–113. AAAI/MIT Press, 1996.

[Ganascia, 1991]J.-G. Ganascia. Charade : apprentissage de bases de connaissances. In Y. Kodratoff and E. Diday, editors, *Induction symbolique et numérique à partir de données*, pages 309–326. Cépaduès Editions, 1991.

[Gras *et al.*, 2004]Régis Gras, Raphaël Couturier, Julien Blanchard, Henri Briand, Pascale Kuntz, and Philippe Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:3–31, 2004. numéro spécial Mesures de qualité pour la fouille de données.

[Gras, 1996]Régis Gras. *L'implication statistique : nouvelle méthode exploratoire de données*. La Pensée Sauvage Editions, 1996.

[Guillet, 2004]Fabrice Guillet. Mesures de la qualité des connaissances en ecd, 2004. Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

[Lallich and Teytaud, 2004]Stéphane Lallich and Olivier Teytaud. Evaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:193–218, 2004. numéro spécial Mesures de qualité pour la fouille de données.

[Lenca *et al.*, 2004]Philippe Lenca, Patrick Meyer, Benoît Vaillant, Philippe Picouet, and Stéphane Lallich. Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1:219–246, 2004. numéro spécial Mesures de qualité pour la fouille de données.

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

[Liu *et al.*, 2000]Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.

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

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

[Silberschatz and Tuzhilin, 1996]Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[Tan *et al.*, 2004]Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.