A Robustness Measure of Association Rules

Yannick Le Bras[1,3], Patrick Meyer[1,3], Philippe Lenca[1,3], and Stéphane Lallich[2]

[1] Institut Télécom, Télécom Bretagne,
UMR CNRS 3192 Lab-STICC,
Technopôle Brest Iroise CS 83818, 29238 Brest Cedex 3
{yannick.lebras,patrick.meyer,philippe.lenca}@telecom-bretagne.eu
[2] Université de Lyon
Laboratoire ERIC, Lyon 2, France
stephane.lallich@univ-lyon2.fr
[3] Université européenne de Bretagne, France

Abstract. We propose a formal definition of the robustness of association rules for interestingness measures. It is a central concept in the evaluation of the rules and has only been studied unsatisfactorily up to now. It is crucial because a good rule (according to a given quality measure) might turn out as a very fragile rule with respect to small variations in the data. The robustness measure that we propose here is based on a model we proposed in a previous work. It depends on the selected quality measure, the value taken by the rule and the minimal acceptance threshold chosen by the user. We present a few properties of this robustness, detail its use in practice and show the outcomes of various experiments. Furthermore, we compare our results to classical tools of statistical analysis of association rules. All in all, we present a new perspective on the evaluation of association rules.

Keywords: association rules, robustness, measure, interest.

1 Introduction

Since their seminal definition [1] and the APRIORI algorithm [2], association rules have generated a lot of research activities around algorithmic issues. Unfortunately, the numerous deterministic and efficient algorithms inspired by APRIORI tend to produce a huge number of rules. A widespread method to evaluate the interestingness of association rules consists of the quantification of this interest through objective quality measures on the basis of the contingency table of the rules. However, the provided rankings may strongly differ with respect to the chosen measure [3]. The large number of measures and their several properties have given rise to many research activities. We suggest that the interested reader refers to the following surveys: [4], [5], [6], [7] and [8].

Let us recall that an association rule $\mathtt{A} \rightarrow \mathtt{B}$, extracted from a database $\mathcal{B}$, is considered as an interesting rule according to the measure m and the user-specified threshold $m_{\min}$, if $m(\mathtt{A} \rightarrow \mathtt{B}) \geq m_{\min}$. This qualification of the rules raises some legitimate questions: to what extent is a good rule the result of

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part II, LNAI 6322, pp. 227–242, 2010.

chance; is its evaluation significantly above the threshold; would it still be valid if the data had been different to some extent (noise) or if the acceptance threshold had been slightly raised; are there interesting rules which have been filtered out because of a threshold which is somewhat too high.

These questions lead very naturally to the intuitive notion of robustness of an association rule, i.e., the sensibility of the evaluation of its interestingness with respect to modifications of $\mathcal{B}$ and/or $m_{\min}$. Besides, it is already obvious here and now that this concept is closely related to the addition of counterexamples and/or the loss of examples of the rule. In this perspective, the study of the measures according to the number of such counterexamples becomes crucial: their decrease according to the number of counterexamples is a necessary condition for their eligibility, whereas their more or less high decrease rate when the first counterexamples appear is a property depending on the user's goals. We recommend that the interested reader has a look at [7] for a detailed study of 20 measures on these two characteristics.

To our knowledge, only very few works concentrate on the robustness of association rules, and can roughly be divided into three approaches: the first one is experimental and is mainly based on simulations [9,10,11], the second one uses statistical tests [5,12], whereas the third one is more formal as it studies the derivative of the measures [13,14,15].

Our proposal, which develops the ideas presented in [13] and [14], gives on the one hand a precise definition of the notion of robustness, and on the other hand presents a formal and coherent measure of the robustness of association rules. In Section 2 we briefly recall some general notions on association rules before presenting the definition of the measure of robustness and its use in practice. Then, in Section 3, we detail some experiments on classical databases with this notion of robustness. We then compare this concept to that of statistical significance in Section 4 and conclude in Section 5.

2 Robustness

2.1 Association Rules and Quality Measures

In a previous work [16], we have focused on a formal framework to study association rules and quality measures, which was initiated by [17]. Our main result in that article is the combination of an association rule with a projection in the unit cube of $\mathbb{R}^3$. As the approach detailed in this article is based on this framework, we briefly recall it here. Let us note $r : \mathtt{A} \rightarrow \mathtt{B}$ an association rule in a database $\mathcal{B}$. A quality measure is a function which associates a rule with a real number characterizing its interest. In this article, we focus exclusively on objective measures, whose value on r is determined solely by the contingency table of the rule. Figure 1 presents such a contingency table, in which we write p_x for the frequency of the pattern $\mathtt{X}$.

Once the three degrees of freedom of the contingency table are chosen, it is possible to consider a measure as a function from $\mathbb{R}^3$ to $\mathbb{R}$ and to use the classical results and techniques from mathematical analysis. In our previous work [16], we

	B	$\overline{\text{B}}$	
A	p_{ab}	$p_{a\bar{b}}$	p_a
$\overline{\text{A}}$	$p_{\bar{a}b}$	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}}$
	p_b	$p_{\bar{b}}$	1

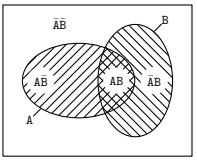


Fig. 1. Contingency table of $r : \mathtt{A} \rightarrow \mathtt{B}$

have shown that it is possible to make a link between algorithmic and analytical properties of certain measures, in particular those related to their variations.

In order to study the measures as functions of three variables, it is necessary to thoroughly define their domain of definition. This domain depends on the chosen parametrization: via the examples, the counterexamples or even the confidence. [14], [7] have stressed out the importance of the number of counterexamples in the evaluation of the interestingness of an association rule. As a consequence, in this work we analyze the behavior of the measures according to the variations of the counterexamples, i.e., an association rule $r : \mathtt{A} \rightarrow \mathtt{B}$ is characterized via the triplet $(p_{a\bar{b}}, p_a, p_b)$. In this configuration, the interestingness measures are functions on a subset $\mathcal{D}$ of the unit cube of $\mathbb{R}^3$ whose definition is given hereafter [16]:

$$\mathcal{D} = \left\{ (x, y, z) \;\middle|\; \begin{array}{ccc} 0 & < y < & 1 \\ 0 & < z < & 1 \\ \max(0, y - z) & < x < & \min(y, 1 - z) \end{array} \right\}$$

where x (resp. y, z) represent $p_{a\bar{b}}$ (resp. p_a, p_b).

In $\mathbb{R}^3$ a rule can now be considered as a vector, and it is possible to study its neighborhood and to observe its behavior in this neighborhood. This represents the starting point for the new characterization of the robustness of association rules, which we introduce in the following section.

2.2 A Definition of the Robustness

Let us suppose that a user wishes to evaluate association rules extracted from a database $\mathcal{B}$ via an objective interestingness measure m. In such a case, he has fixed a threshold $m_{\min}$ above which the rules are considered as interesting. These selected rules depend on many parameters, among which:

- the threshold $m_{\min}$: the user can modify it at any time and let appear or disappear a large number of rules;
- the noise: a given selected rule might not resist variations of the data, as, e.g., the addition of new transactions or the presence of erroneous recordings.

In this article we propose a contribution to the study of this latter point, namely the weakness of a rule according to variations in the data. [14] suggest different approaches for the study of the variations of the measures according to counterexamples of the rules. They develop various models to study the variations in

the data that a rule can withstand in order to remain interesting. However, the authors do not give a general model which aggregates their multiple proposals, which does not allow to obtain a general measure of the robustness.

Our vision of the robustness is quite different and is based on the concept of *limit rule*. Note right beforehand that such a rule can be abstract, as it is not necessarily a rule which is achieved in the database $\mathcal{B}$. We define a distance between two rules r and r', $d_2(r, r')$, which is the euclidian distance between the projection of r and r' in $\mathcal{D}$.

Definition 1 (Limit rule). *A limit rule is an association rule $r_{\min}$, possibly abstract, such that $m(r_{\min}) = m_{\min}$. Let r be an association rule. We write r^* for a limit rule which minimizes $d(r, r_{\min})$ in $\mathbb{R}^3$. Formally,*

$$r^* \in argmin\{d_2(r, r_{\min}) | r_{\min} \text{ limit rule}\}$$

The limit rules which are actually realized in the database are those rules which have been barely selected according to the threshold $m_{\min}$. For a given rule r, r^* is not necessarily unique. However, its choice is not crucial for the notion of robustness that we are introducing in the sequel.

As a limit rule is an association rule, associated with $(x_{\min}, y_{\min}, z_{\min})$, it is necessarily an element of $\mathcal{D}$. Therefore, $d(r, r^*)$ is not simply the distance between r and the surface $\mathcal{S}$ of equation $m = m_{\min}$, but rather the distance to $\mathcal{S} \cap \mathcal{D}$.

Definition 2 (Robustness of an association rule). *Let m be an interestingness measure and $m_{\min}$ a threshold fixed by the user. Let r be an association rule on a database $\mathcal{B}$ such that $m(r) \geq m_{\min}$. The robustness of r according to m and $m_{\min}$ is defined by:*

$$rob_m(r, m_{\min}) = \frac{d(r, r^*)}{\sqrt{3}}$$

Figure 2 shows our concept of robustness for two rules. The important factor in this formula is the numerator $d(r, r^*)$, the division by $\sqrt{3}$ is a normalization factor which allows to fit the quantity in the interval $[0, 1]$. Other normalizations are indeed possible. If there is no ambiguity, we will write this robustness $\text{rob}(r)$. In the following section we discuss this definition to show why it represents a notion of robustness, and present some of its properties.

2.3 Properties of the Robustness

Let us start by justifying the designation of robustness. Consider a database $\mathcal{B}$ and an association rule $r : \mathtt{A} \rightarrow \mathtt{B}$ in $\mathcal{B}$ such that $m(r) > m_{\min}$. We note $(p_{a\bar{b}}, p_a, p_b)$ the corresponding supports. Let us now add some noise in the database $\mathcal{B}$ in order to obtain a database $\mathcal{B}'$ in which the rule $r' : \mathtt{A} \rightarrow \mathtt{B}$ is characterized by $(p'_{a\bar{b}}, p'_a, p'_b)$. For short, after the noise introduction the patterns remain the same, but their supports change. Let us now suppose that the noise which is added respects:

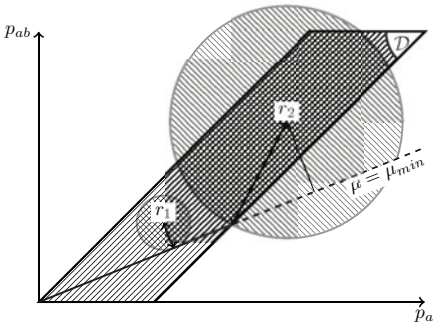


Fig. 2. Visualization of robustness for two different rules r_1 and r_2. Here, p_b is fixed, and the measure is the measure of confidence.

$$|p'_{a\bar{b}} - p_{a\bar{b}}| \leq \frac{d(r, r^*)}{\sqrt{3}} \ ; \ |p'_a - p_a| \leq \frac{d(r, r^*)}{\sqrt{3}} \ ; \ |p'_b - p_b| \leq \frac{d(r, r^*)}{\sqrt{3}}$$

In such a case, $d(r, r') = \sqrt{|p'_{a\bar{b}} - p_{a\bar{b}}|^2 + |p'_a - p_a|^2 + |p'_b - p_b|^2} \leq d(r, r^*)$, and thus by the definition of r^*, $m(r') \geq m_{\min}$. Thus, $\text{rob}(r)$ clearly expresses the quantity of noise that the rule can withstand and still stay interesting. We can see that our definition of the robustness is closely linked to a notion of safety: if the noise is sufficiently controlled, then an interesting rule will stay interesting. The inverse is however not true, as a poorly robust rule can evolve to become more interesting and more robust.

This notion of robustness can be easily understood if the noise is inserted *by transaction*. Indeed, if one inserts the noise into less than $\text{rob}(r)\%$ of the transactions, the rule r will stay interesting according to $m_{\min}$. However, if the noise is inserted *by attribute* [9], it is harder to control it accurately.

Inversely, if the percentage of noise in a database is known, then the interesting robust rules (for this amount of noise) extracted from the noisy database will also be interesting in the *ideal* noiseless one.

Property 1. The robustness measure $rob(r)$ has the following interesting analytical characteristics :

- the robustness of a rule is a real number of $[0, 1]$;
- $\text{rob}_m(r, m_{\min}) = 0$ if r is a limit rule, i.e., if $m(r) = m_{\min}$;[1]
- if the measure m, seen as a function of 3 variables, is continuous from $\mathcal{D} \subset \mathbb{R}^3$ to $\mathbb{R}$, then the robustness is decreasing with respect to $m_{\min}$;
- the robustness is continuous according to r.

[1] Note that the value $\text{rob}_m(r, m_{\min}) = 1$ is a theoretical value which corresponds to a very special configuration of r, $m_{\min}$ and m. In practice, in our experiments, we have not encountered this value.

These properties allow us to confirm certain expected behaviors of the robustness notion. First, the higher the threshold is, the less robust are the rules, and the more important is the reliability of the data. Second, two rules having close projections in $\mathbb{R}^3$ will have equivalent values for the robustness.

2.4 Calculating the Robustness

The calculation of the robustness requires the determination of the distance to a surface under certain constraints. For complex measures (Klosgen, collective strength, ...), this calculation cannot be performed in a formal way, and necessitates numerical techniques. However, there exist a certain number of measures based on frequencies for which the calculation is quite simple. In this paper we concentrate exclusively on these measures, which we call *planar measures*.

Definition 3 (Planar measure). *An interestingness measure m is called* planar *if the surface defined by $m(r) = m_{\min}$ is a plane.*

In particular, this is the case for measures like Sebag-Shoenauer, example-counterexample rate, Jaccard, contramin, precision, recall, specificity. In this case, the distance between a rule r_1 with coordinates (x_1, y_1, z_1) and the plane $\mathcal{P} : ax + by + cz + d = 0$ is given by:

$$d(r_1, \mathcal{P}) = \frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}}$$

However, to obtain the robustness measure, r^* must belong to the domain $\mathcal{D}$. Therefore, if it is not the case for the orthogonal projection of the rule on the plane, the distance of interest is the one between the rule and the intersection polygon $\mathcal{P} \cap \mathcal{D}$. We therefore determine the corners of this convex polygon to obtain the distance between the rule and the perimeter of the polygon as the minimal distance between the rule and the edges of the polygon (as segments).

Consequently, the calculation algorithm of the robustness measure for planar measures is given hereafter:

- Determine $r^{\perp}$, the orthogonal projection of r on $\mathcal{P}$;
- If $r^{\perp} \in \mathcal{D}$, $r^* = r^{\perp}$ and return $d(r, r^*)$;
- Else, return the distance between the rule and the perimeter of the intersection polygon.

Example 1. The following measures are planar. Their level lines $m = m_0$ define the following planes:

- confidence: $x - (1 - m_0)y = 0$;
- Sebag-Shoenauer: $(1 + m_0)x - y = 0$;
- example-counterexample rate: $(2 - m_0)x - (1 - m_0)y = 0$;
- Jaccard: $(1 + m_0)x - y + m_0 z = 0$.

Let us now study in further details the case of the confidence measure. In a parametrization via the counterexamples, the plane defined by the confidence threshold $m_{\min}$ is $\mathcal{P} : x - (1 - m_{\min})y = 0$. The distance between a rule r_1 (with coordinates (x_1, y_1, z_1) and confidence $m(r_1) > m_{\min}$) and the plane is given by

$$d = y_1 \frac{m(r_1) - m_{\min}}{\sqrt{1 + (1 - m_{\min})^2}}. \quad (1)$$

Thereby, for a given value of $m_{\min}$, the robustness depends on two parameters:

- y_1, the support of the antecedent;
- $m(r_1)$, the value taken by the interestingness measure of the rule.

Thus, two rules having the same confidence, can have very different robustness values. Similarly, two rules having the same robustness, can have various confidences. Therefore, it will not be surprising to observe rules with a low value for the interestingness measure and a high robustness, as well as rules with a high interestingness and a low robustness. Indeed, it is possible to discover rules which are simultaneously very interesting and very fragile.

Example 2. Consider a fictive database of 100000 transactions. We write n_x for the number of occurrences of the pattern `X`. In this database, we can find a first rule $r_1 : \texttt{A} \rightarrow \texttt{B}$ such that $n_a = 100$ and $n_{a\bar{b}} = 1$. Its confidence equals 99%. However, its robustness, at the level of confidence of 0.8 equals $\text{rob}(r_1) = 0.0002$. A second rule $r_2 : \texttt{C} \rightarrow \texttt{D}$ has the following characteristics: $n_c = 50000$ and $n_{c\bar{d}} = 5000$. Its confidence only equals 90%, whereas its robustness measure is 0.05. As r_2 has proportionally to its antecedent more counterexamples than r_1, at first sight it could be mistakenly considered as less reliable.

In the first case, the closest limit rule can be described by $n^*_a = 96$ et $n^*_{a\bar{b}} = 19$. The original rule therefore only resists very few variations on the entries. The second rule however has a closest limit rule with parameters $n_c = 49020$ et $n_{c\bar{d}} = 9902$, which shows that r_2 can bear about a thousand changes in the database.

As a conclusion, r_2 is much less sensitive to noise as r_1, even if r_1 appears to be more interesting according to the confidence measure.

These observations show that the determination of the *real interestingness* of a rule is more difficult than it seems: how should we arbitrate between a rule which is interesting according to a quality measure but poorly robust, and one which is less interesting but which is more reliable with respect to noise.

2.5 Use of the Robustness in Practice

The robustness, as defined earlier, can have two immediate applications. First, the robustness measure allows to compare any two rules and to compute a weak order on the set of selected rules (a ranking with ties). Second, the robustness measure can be used to filter the rules if the user fixes a limit threshold.

However, similarly as for the interestingness measures, the determination of this robustness threshold might be a difficult task. In practice, it should therefore be avoided to impose the determination of another threshold on a user. This notion can nevertheless be a further parameter in the comparison of two rules.

When considering the interestingness measure of a rule according to its robustness measure, it is possible to distinguish between two situations. When comparing rules which are fragile and uninteresting to rules which are robust and interesting, it is obvious that a user will prefer the second ones. However, this choice is more demanding for a fragile but interesting rule compared to a robust but uninteresting one. Is it better to have an interesting rule which depends a lot on the noise in the data or a very robust one, which will resist changes in the data? The answers to this question depend of course on the practical situation and the confidence of the user in the quality of his data.

In the sequel we will observe that the interestingness vs. robustness plots show a lot of robust rules which are dominated in terms of quality measures by less robust ones.

3 Experiments

In this section we present the results obtained on 4 databases for 5 planar measures. First we present the experimental protocol, then we study the plots that we generated in order to stress out the link between the interestingness measures and the robustness. Finally, we analyze the influence of noise on association rules.

3.1 Experimental Protocol

Extraction of the rules. Recall that we focus here on planar measures. For this experiment, we have selected 5 of them: confidence, Jaccard, Sebag-Shoenauer, example-counterexample rate, and specificity. Table 1 summarizes their definition in terms of the counterexamples and the plane they define in $\mathbb{R}^3$.

For our experiments we have chosen 4 of the usual databases [18]. We have extracted class rules, i.e. rules for which the consequent is constrained, both from Mushroom and a discretized version of Census. The databases Chess and Connect have been binarized in order to extract unconstrained rules. All the rules

Table 1. The planar measures, their definition, the plane defined by m_0 and the selected threshold value

name	formula	plane	threshold
confidence	$\frac{p_a - p_{a\bar{b}}}{p_a}$	$x - (1 - m_0)y = 0$	0.984
Jaccard	$\frac{p_a - p_{a\bar{b}}}{p_b + p_{a\bar{b}}}$	$(1 + m_0)x - y + m_0 z = 0$	0.05
Sebag-Shoenauer	$\frac{p_a - p_{a\bar{b}}}{p_{a\bar{b}}}$	$(1 + m_0)x - y = 0$	10
specificity	$\frac{1 - p_b - p_{a\bar{b}}}{1 - p_a}$	$x - m_0 y + z = 1 - m_0$	0.5
example-counterexample rate	$1 - \frac{p_{a\bar{b}}}{p_a - p_{a\bar{b}}}$	$(2 - m_0)x - (1 - m_0)y = 0$	0.95

Table 2. Databases used in our experiments. The fifth column indicates the maximal size of the extracted rules.

database	attributes	transactions	type	size	# rules
census	137	48842	class	5	244487
chess	75	3196	unconstrained	3	56636
connect	129	67557	unconstrained	3	207703
mushroom	119	8124	class	4	42057

have been generated via the APRIORI algorithm of [19], in order to obtain rules with a positive support, a confidence above 0.8 and of variable size according to the database. These information are summarized in Table 2. Note that the generated rules are interesting, the nuggets of knowledge have not been left out, and the number of rules is fairly high.

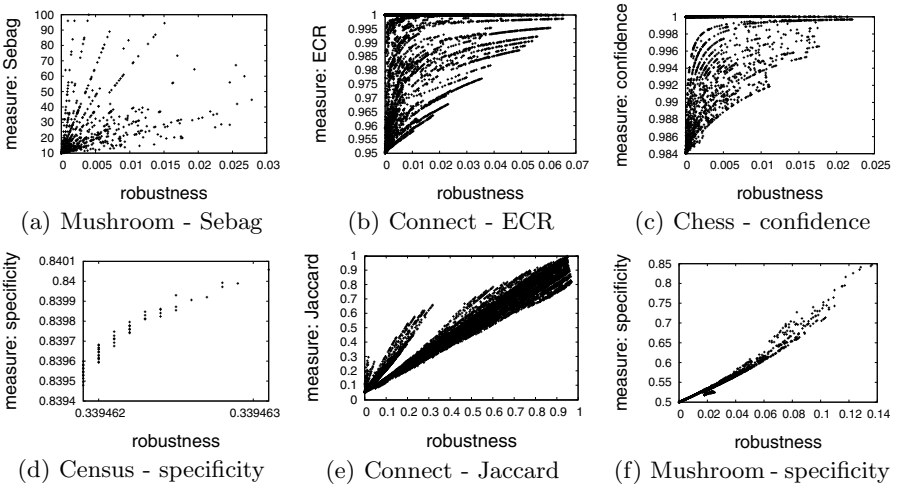
Calculation of the robustness. For each set of rules and each measure we have applied the same calculation method for the robustness of the association rules. In a first step, we have selected only the rules with an interestingness measure above a predefined threshold. We have chosen to fix this threshold according to the values of Table 1. These thresholds have been determined by observing the behavior of the measures on the rules extracted from the Mushroom database, in order to obtain interesting and uninteresting rules in similar proportions.

Then we have implemented an algorithm, based on the description of Section 2.4 for the specific case of planar measures, which determines the robustness of a rule according to the value it takes for the interestingness measure and the threshold. As an output we obtain a list of rules with their corresponding support, robustness and interestingness measure values. The complexity of this algorithm depends mostly on the number of rules which have to be analyzed. These results, presented in Section 3.2, allow us to generate the interestingness vs. robustness plots mentioned earlier.

Noise insertion. As indicated earlier, we analyze the influence of noise in the data on the rules, according to their robustness. This noise is introduced transaction-wise, for the reasons mentioned in Section 2.3, as follows: in 5% of randomly selected rows of each database, the values of the attributes are modified randomly (equally likely and without replacement). Once the noise is inserted, we recalculate the supports of the initially generated rules. We then extract the interesting rules according to the given measures and evaluate their robustness. The study of the noise is discussed in Section 3.3.

3.2 Robustness Analysis

For each database and each interestingness measure, we plot the value taken by the rule for the measure according to its robustness. Figure 3 shows a representative sample of these results (for a given interestingness measure, the plots are, in general, quite similar for all the databases).



(a) Mushroom - Sebag (b) Connect - ECR (c) Chess - confidence

(d) Census - specificity (e) Connect - Jaccard (f) Mushroom - specificity

Fig. 3. Value of the interestingness measure according to the robustness for a sample of databases and measures

Various observations can be deduced from these plots. First, the interestingness measure is in general increasing with the robustness. A closer analysis shows that a large number of rules are dominated in terms of their interestingness by less robust rules. This is specially the case for the Sebag measure (Figure 3(a)), for which we observe that a very interesting rule r_1 ($\text{Sebag}(r_1) = 100$) can be significantly less robust ($\text{rob}(r_1) = 10^{-4}$) than a less interesting rule r_2 ($\text{Sebag}(r_2) = 20$ and $\text{rob}(r_2) = 2 \cdot 10^{-3}$). The second rule resists twenty times more changes than the first one.

Second, in most of the cases, we observe quite distinct level lines. Sebag and Jaccard bring forward straight lines, confidence and example-counterexample rate generate concave curves, and the specificity seems to produce convex ones.

Let us analyze the case of the level curves for the confidence. Note that similar calculations can be done for the other interestingness measures. Equation (1) presents the robustness according to the measure, where y represents p_a. As $p_a = \frac{p_{a\bar{b}}}{1-m(r)}$, we can write the measure $m(r)$ according to the distance d:

$$m(r) = \frac{m_{\min} + \sqrt{1 + (1 - m_{\min})^2} \cdot \frac{d}{x}}{1 + \sqrt{1 + (1 - m_{\min})^2} \cdot \frac{d}{x}} \tag{2}$$

Thus, for a given x (i.e. for a constant number of counterexamples), the rules are situated on a well defined concave and increasing curve. This shows that the level lines in the case of the confidence are made of rules which have the same number of counterexamples.

Another behavior seems common for most of the studied measures: there exists no rule which is close to the threshold and very robust. Sebag is the only

measure which does not completely fit to this observation. We think that this might be strongly linked to the restriction of the study to planar measures.

3.3 Study of the Influence of the Noise

In this section we are studying the links between the addition of noise to the database and the evolution of the rules sets, with respect to robustness. To do so, we create 5 noisy databases from the original ones (see 3.1) and for each of them, analyze the robustness of the rules resisting these changes and of the ones disappearing. In order to validate our notion of robustness, we expect that the robustness of the rules which have vanished is lower on average than the robustness of the rules which stay in the noisy databases.

Table 3 presents the results of this experiment, by showing the average of the robustness values for the two sets of rules, for the 5 noisy databases. In most of the cases, the rules which resisted the noise are approximatively 10 times more robust than those which vanished. The only exception is the Census database for the measures example-counterexample rate, Sebag and confidence, which do not confirm this result. However, this is not negating our theory. Indeed, the initial robustness values for the rules of the Census database are around 10^{-6}, which makes them vulnerable to 5% of noise. It is therefore not surprising that all the rules can potentially become uninteresting.

On the opposite, the measure of specificity underlines a common behavior of the Census and the Connect databases. For both of them, no rule vanishes after the insertion of 5% of noise. The average value of the robustness of the rules which resisted the noise is significantly higher than these 5%, which means that all the rules are well protected. In the case of the Census base, the lowest specificity value equals 0.839, which is well above the threshold which has been fixed beforehand. This explains why the rules originating from the Census database

Table 3. Comparison between the average robustness values for the vanished rules and those which resisted the noise, for each of the studied measures

(a) example-counterexample rate

base	vanished	stayed
census	0.83e-6	0.79e-6
chess	1.16e-3	0.96e-2
connect	5.26e-4	7.72e-3
mushroom	9.4e-5	6.6e-4

(b) Sebag

base	vanished	stayed
census	1.53e-6	1.53e-6
chess	1.63e-3	1.72e-2
connect	8.38e-4	1.42e-2
mushroom	1.28e-4	1.22e-3

(c) specificity

base	vanished	stayed
census	0	0.19
chess	7.23e-5	8.76e-2
connect	0	1.2e-1
mushroom	2.85e-4	1.37e-2

(d) confidence

base	vanished	stayed
census	2.61e-7	2.61e-7
chess	5.59e-4	3.77e-3
connect	2.16e-4	2.73e-3
mushroom	5.51e-5	2.34e-4

(e) Jaccard

base	vanished	stayed
census	0	0
chess	3.2e-4	1.69e-1
connect	1.94e-3	1.43e-1
mushroom	3.20e-4	1.90e-2

all resist the noise. In the case of the Connect database, the average value of the specificity measure equals 0.73 with a standard deviation of 0.02. The minimal value equals 0.50013 and corresponds to a robustness of $2.31e-5$. However, this rule has been saved in the 5 noise additions. This underlines the fact that our definition of the robustness corresponds to the definition of a security zone around a rule. If the rule changes and leaves this area, it can evolve freely in the space, without ever getting to the threshold surface. Nevertheless, the risk still prevails.

In the following section we compare the approach via the robustness measure to a more classical one to determine if a rule is considered as statistically significant.

4 Robustness vs. Statistical Significance

In the previous sections, we have defined the robustness of a rule as its capacity to overcome variations in the data, like a loss of examples and / or a gain of counter-examples, so that its evaluation $m(r)$ remains above the given threshold $m_{\min}$. This definition looks quite similar to the notion of statistical significance. In this section we explore the links between both approaches.

4.1 Significant Rule

From a statistical point of view, we have to distinguish between the following notions: $m(r)$ is the empirical value of the rules computed over a given data sample, that is the observed value of the random variable $M(r)$, and $\mu(r)$ is the theoretical value of the interestingness measure. A *statistically significant rule* r for a threshold $m_{\min}$ and the chosen measure is a rule for which we can consider that $\mu(r) > m_{\min}$.

Usually, for each rule, the null-hypothesis $H_0 : \mu(r) = m_{\min}$ is tested against the alternative hypothesis $H_1 : \mu(r) > m_{\min}$. A rule r is considered as significant at the significance level α_0 (type I error, false positive) if its p-value is at most α_0. Recall that the p-value of a rule r whose empirical value is $m(r)$ is defined as $P(M(r) \geq m(r)|H_0)$.

However, due to the high number of tests which need to be performed, and the resulting multitude of false discoveries, the p-values need to be adapted (see [20] for a general presentation, and [5] for the specific case of association rules with respect to independency).

The algebraic form of the p-value can be determined only if the law of M under H_0 is (at least approximately) known. This is the case for the measure of confidence, for which $M = N_{ab}/N_a$ where N_x is the number of instances of the itemset x. The distribution of M under H_0 is established via the models proposed by [21] and generalized by [22], provided that the margins N_a and N_b are fixed. However, this is somewhat simplistic, like for the χ^2 test. Furthermore, in many cases, as e.g. for the planar measure of Jaccard, it is impossible to establish the law of M under H_0.

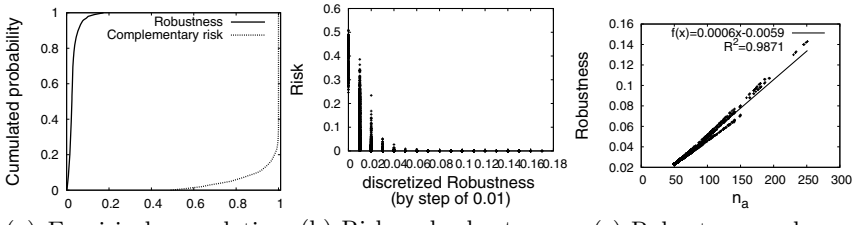
Therefore, we here prefer to estimate the risk that the interestingness measure of the rule falls below the threshold $m_{\min}$ via a bootstrapping technique which allows to approximate the variations of the rule in the real population. In our case we draw with replacement 400 samples of size n from the original population of size n. The risk is then estimated via the proportion of samples in which the evaluation of the rule fell under the threshold. Note that this value is smoothed by using the normal law. Only the rules with a risk less or equal to α_0 are considered as significant.

However, even if no rule is significant, $n\alpha_0$ rules will be selected. In the case where $n = 10000$ and $\alpha_0 = 0.05$, this would lead to 500 false discoveries. Among all the false discoveries control methods, one is of particular interest. In [23], Benjamini and Liu proposed a sequential method: the risk values are sorted in increasing order and named $p_{(i)}$. A rule is selected if its corresponding $p_{(i)} \leq i\frac{\alpha_0}{n}$. This procedure allows to control the expected proportion of wrongfully selected rules in the set of selected rules (False Discovery Rate) conditionally to the independence of the data. This is compatible with positively dependent data.

4.2 Comparison of the Two Approaches on an Example

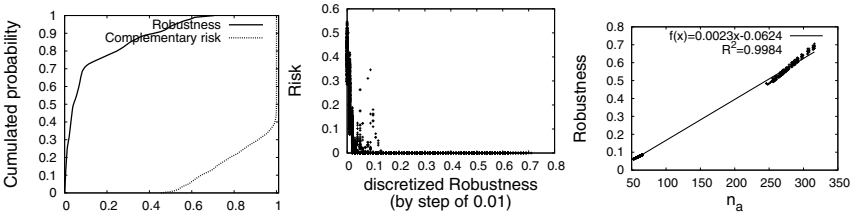
In order to get a better understanding of the difference between these rules stability approaches, we compare the results of the robustness calculation and the complementary risk resulting from the bootstrapping. Our experiments are based on the SolarFlare database [18]. We detail here the case of the two measures mentioned above, Confidence and Jaccard, for which an algebraic approach of the p-value is either problematic (fixed margins), or impossible.

We first extract the rules with the classical Apriori algorithm with support and confidence thresholds set to 0.13 and 0.85 respectively. This produces 9890 rules with their confidence and robustness. A bootstrapping technique with 400 iterations allows to compute the risk of each rule to fall below the threshold. From the 9890 rules, one should note that even if 8481 have a bootstrapping risk of less than 5%, only 8373 of them are kept when applying the procedure of Benjamini and Liu.



(a) Empirical cumulative distribution functions (b) Risk and robustness (c) Robustness and n_a

Fig. 4. Confidence Case



(a) Empirical cumulative distribution functions (b) Risk and robustness (c) Robustness and n_a

Fig. 5. Jaccard index case

Figure 4(a) shows the empirical cumulative distribution function of the robustness and the complementary risk resulting from the bootstrapping. It shows that the robustness is clearly more discriminatory than the complementary risk, especially for interesting rules. Figure 4(b) represents the risk with regard to the class of robustness (discretized by steps of 0.01). It shows that the risk is globally correlated with robustness.

However, the outputs of two approaches are clearly different. On the one side, the process of Benjamini returns 1573 unsignificant rules having a robustness less than 0.025 (except for 3 of them). On the other side, 3616 rules of the significant ones have a robustness less than 0.05. Besides, it is worth noticing that the robustness of the 2773 logical rules takes many different values between 0.023 and 0.143. Finally, as shown in Figure 4(c), the robustness of a rule is linearly correlated with its coverage.

The results obtained with the Jaccard measure are of the same kind. The support threshold is set to 0.0835, whereas the Jaccard index is fixed to 0.3. We obtain 6066 rules, from which 4059 are declared significant at the 5% level by the bootstrapping technique (400 iterations), and 3933 by the process of Benjamini (that is 2133 unsignificant rules). Once again, the study of the empirical cumulative distribution functions (see Figure 5(a)) shows that the robustness is more discriminatory than the complementary risk of the bootstrapping for the more interesting rules. Similarly, Figure 5(b) shows that the risk for the Jaccard measure is globally correlated with the robustness, but again, there are significant differences between the two approaches. The rules determined as significant for the process of Benjamini have a robustness less than 0.118 when significant rules at the 5% level have robustness spread from 0.018 and 0.705, which is a quite big range.

There are 533 rules with a Jaccard index greater than 0.8. All of them have a zero complementary risk, and their robustness value vary between 0.062 and 0.705. As shown by Figure 5(c), the robustness of the Jaccard index is linearly correlated to the coverage of the rule for high values of the index (> 0.80).

As a conclusion of this comparison, the statistical approach of bootstrapping to estimate the type I error has the major drawback that it is not very discriminatory, especially for high values of n, which is the case in datamining.

In addition, the statistical analysis assume that the actual data are a random subset of the whole population, which is not really the case in datamining. All in all, the robustness study for a given measure gives a more precise idea of the stability of interesting rules.

5 Conclusion

The robustness of association rules is a crucial topic, which has only been poorly studied by formal approaches. The robustness of a rule with respect to variations in the database adds a further argument for its interestingness and increases the validity of the information which is given to the user.

In this article, we have presented a new operational notion of robustness which depends on the chosen interestingness measure and the corresponding acceptability threshold. As we have shown, our definition of this notion is consistent with the natural intuition linked to the concept of robustness. We have analyzed the case of a subset of measures, called planar measures, for which we are able to give a formal characterization of the robustness. Our experiments on 5 measures and 4 classical databases illustrate and corroborate the theoretical discourse. The proposed robustness measure is also compared to a more classical statistical analysis of the significance of a rule, which turns out to be less discriminatory in the context of data mining.

In practice, the robustness measure allows to rank rules according to their ability to withstand changes in the data. However, the determination of a robustness threshold by a user remains an issue. In the future, we plan to propose a generic protocol to calculate the robustness of association rules with respect to any interestingness measure via the use of numerical methods.

References

1. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, pp. 478–499 (1994)
3. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. In: 7th International Conference on Discovery Science, Padova, Italy, pp. 290–297 (2004)
4. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3, Article 9) (2006)
5. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: Measure and statistical validation. In: Quality Measures in Data Mining, pp. 251–275 (2007)
6. Geng, L., Hamilton, H.J.: Choosing the right lens: Finding what is interesting in data mining. Quality Measures in Data Mining, 3–24 (2007)
7. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. European Journal of Operational Research 184, 610–626 (2008)

8. Suzuki, E.: Pitfalls for categorizations of objective interestingness measures for rule discovery. In: Statistical Implicative Analysis, Theory and Applications, pp. 383–395 (2008)
9. Azé, J., Kodratoff, Y.: Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In: 2nd Extraction et Gestion des Connaissances conference, Montpellier, France, pp. 143–154 (2002)
10. Cadot, M.: A simulation technique for extracting robust association rules. In: Computational Statistics & Data Analysis, Limassol, Chypre (2005)
11. Azé, J., Lenca, P., Lallich, S., Vaillant, B.: A study of the robustness of association rules. In: The 2007 Intl. Conf. on Data Mining, Las Vegas, Nevada, USA, pp. 163–169 (2007)
12. Rakotomalala, R., Morineau, A.: The TVpercent principle for the counterexamples statistic. In: Statistical Implicative Analysis, Theory and Applications, pp. 449–462. Springer, Heidelberg (2008)
13. Lenca, P., Lallich, S., Vaillant, B.: On the robustness of association rules. In: 2nd IEEE International Conference on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics, Bangkok, Thailand, pp. 596–601 (2006)
14. Vaillant, B., Lallich, S., Lenca, P.: Modeling of the counter-examples and association rules interestingness measures behavior. In: The 2006 Intl. Conf. on Data Mining, Las Vegas, Nevada, USA, pp. 132–137 (2006)
15. Gras, R., David, J., Guillet, F., Briand, H.: Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association. In: 3rd Workshop on Qualite des Donnees et des Connaissances, Namur Belgium, pp. 35–43 (2007)
16. Le Bras, Y., Lenca, P., Lallich, S.: On optimal rules discovery: a framework and a necessary and sufficient condition of antimonotonicity. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 705–712. Springer, Heidelberg (2009)
17. Hébert, C., Crémilleux, B.: A unified view of objective interestingness measures. In: 5th Intl. Conf. on Machine Learning and Data Mining, Leipzig, Germany, pp. 533–547 (2007)
18. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
19. Borgelt, C., Kruse, R.: Induction of association rules: Apriori implementation. In: 15th Conference on Computational Statistics, Berlin, Germany, pp. 395–400 (2002)
20. Dudoit, S., van der Laan, M.J.: Multiple Testing Procedures with Applications to Genomics (2007)
21. Lerman, I.C., Gras, R., Rostam, H.: Elaboration d'un indice d'implication pour les données binaires, I et II. Mathématiques et Sciences Humaines, 5–35, 5–47 (1981)
22. Lallich, S., Vaillant, B., Lenca, P.: A probabilistic framework towards the parameterization of association rule interestingness measures. Methodology and Computing in Applied Probability, 447–463 (2007)
23. Benjamini, Y., Liu, W.: A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference 82(1-2), 163–170 (1999)