Association rule interestingness: measure and statistical validation

Stéphane Lallich, Olivier Teytaud, Elie Prudhomme

Association rule interestingness: measure and statistical validation

Stephane Lallich[1], Olivier Teytaud[2], and Elie Prudhomme[1]

[1] Université Lyon 2, Equipe de Recherche en Ingénierie des Connaissances, 5 Avenue Pierre Mendès-France, 69676 BRON Cedex, France
stephane.lallich@univ-lyon2.fr, eprudhomme@eric.univ-lyon2.fr
[2] TAO-Inria, LRI, CNRS-Université Paris-Sud, bat. 490, 91405 Orsay Cedex, France teytaud@lri.fr

Summary. The search for interesting Boolean association rules is an important topic in knowledge discovery in databases. The set of admissible rules for the selected support and confidence thresholds can easily be extracted by algorithms based on support and confidence, such as *Apriori*. However, they may produce a large number of rules, many of them are uninteresting. One has to resolve a two-tier problem: choosing the measures best suited to the problem at hand, then validating the interesting rules against the selected measures. First, the usual measures suggested in the literature will be reviewed and criteria to appreciate the qualities of these measures will be proposed. Statistical validation of the most interesting rules requests performing a large number of tests. Thus, controlling for false discoveries (type I errors) is of prime importance. An original bootstrap-based validation method is proposed which controls, for a given level, the number of false discoveries. The interest of this method for the selection of interesting association rules will be illustrated by several examples.

Key words: Association rules, interestingness measures, measure properties, multiple testing, false discoveries.

1 Introduction

The association between Boolean variables has been studied for a long time, especially in the context of 2×2 cross-tables. As Hajek and Rauch [21] point out, one of the first methods used to look for association rules is the GUHA method, proposed by Hajek *et al.* [22], where the notions of support and confidence appear. Work done by Agrawal *et al.* [2], Agrawal and Srikant [1], Mannila *et al.* [37] on the extraction of association rules from transactional databases has renewed the interest in the association rules.

In such a database, each record is a transaction (or more generally, a case) whereas the fields are the possible items of a transaction. Let n be the number

of transactions and p the number of items. A Boolean variable is associated to each item. It takes the value "1" for a given transaction if the considered item is present in this transaction, "0" else. The set of transactions form a $n \times p$ Boolean matrix. To each itemset is associated a Boolean variable which is the conjunction of the Boolean variables associated to each item of the considered itemset.

From the Boolean matrix showing which items are the objects of which transaction, one extracts rules like "if a client buys bread and cheese, he is quite likely to also buy wine". A rule of association is an expression $A \rightarrow B$, where A and B are disjoint itemsets. More generally, this form can be applied to any data matrix, as long as continuous variables are discretized and categorical variables are dichotomized.

As the number of possible association rules grows exponentially with the number of items, selecting the "interesting" rules is paramount. Now, one needs to measure how interesting a rule is, and to validate the truly interesting rules with respect to said measure. Previous work done by the authors on the measure [28, 29] and on the validation of the association rules [45, 30] is synthesized in this chapter. Measuring the interest of a rule requires that the user chooses those best adapted to his data and his goal, targeting of group or prediction. Various criteria are presented. Once a measure has been selected and that rules are assessed using that measure, they still must be validated. One could retain the 50 or 100 rules with the highest scores, but these need not be interesting. Whenever possible, one should set a practical or probabilistic threshold. When the measure exceeds the threshold, either the rule is really interesting (true discovery), or it is merely an artefact of the random choice and the rule is not really interesting (false discovery). Each rule must be tested, which mechanically leads to a multitude of false discoveries (or false positives). The authors propose a bootstrap-based method to select interesting rules while controlling the number of false discoveries.

Criteria that can be used to assess measures appropriate to one's goal are presented in Sect. 2. In Sect. 3, it is shown that the validation of rules identified by the selected measures relies on a multitude of tests and the authors propose a multiple test method that controls the number of false discoveries.

2 Measuring Association Rule Interestingness

In this section, we will first look at the support-confidence approach (Sect. 2.1). Then, the notion of rules, implication and equivalences are examined (Sect. 2.2). A list of measures and several assessment criteria are given in the following subsections (Sects. 2.3 and 2.4). In the last subsection, some common features of these measures are highlighted (Sect. 2.5).

2.1 Appeal and Limitations of the Support-Confidence Approach

Support and Confidence

Let n_a and n_b the respective number of A and B transactions, and let n_{ab} be the number of transactions where A and B items appear simultaneously. The support of the rule $A \rightarrow B$ is the proportion of joint A and B transactions:

$SUP\,(A \rightarrow B) = p_{ab} = \frac{n_{ab}}{n}$.

whereas the confidence is the proportion of B transactions among the A transactions, that is the conditional frequency of B given A:

$CONF\,(A \rightarrow B) = \frac{p_{ab}}{p_a} = \frac{n_{ab}}{n_a} = 1 - \frac{n_{a\bar{b}}}{n_a}$.

"Support-Confidence" Extraction Algorithms

Following *Apriori*, the founding algorithm [1], support-confidence extraction algorithms exhaustively seek the association rules, the support and the confidence of which exceed some user-defined thresholds noted min_{SUP} and min_{CONF}. They look for frequent itemsets among the lattice of itemsets, that is, those itemsets whose support exceeds min_{SUP}, using the principle of antimonotonicity of support on the lattice of itemsets:

- any subset of a frequent itemset is frequent
- any superset of a non-frequent itemset is non-frequent.

Then, for each frequent itemset X, the support-confidence algorithms only keep rules of the type $X \backslash Y \rightarrow Y$, with $Y \subset X$, the confidence of which exceeds min_{CONF}.

Pros and Cons of Support-Confidence Approach

The antimonotonicity property of the support makes the support-confidence approach to rule extraction quite appealing. However, its usefulness is questionable, even though the very meaning of support and confidence are translated in easy-to-grasp measures.

First, algorithms of this type generate a very large number of rules, many of them of little interest. Moreover, the support condition, at the core of the extraction process, neglects rules with a small support though some may have a high confidence thus being of genuinely interesting, a common situation in marketing (the so-called nuggets of data mining). If the support threshold is lowered to remedy this inconvenient, even more rules are produced, choking the extraction algorithms.

Finally, the support and confidence conditions alone do not ensure rules with a real interest. Indeed, if the confidence of the rule $A \rightarrow B$ is equal to the marginal frequency of B, namely $p_{b/a} = p_b$, which means that A and B are independent, then the rule $A \rightarrow B$ adds no information (e.g. $p_a = 0.8$, $p_b = 0.9$, $p_{ab} = 0.72$, $p_{b/a} = 0.9$)!

Hence, measures other than support and confidence must be examined, thus promoting some amount of inductive bias.

Table 1. Notations for the joint distribution of itemsets A and B

$A \setminus B$	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	1

2.2 Rule, Implication and Equivalence

Association rules, implication and equivalence must be distinguished; let A and B be two itemsets whose joint distribution is given in Table 1, where 0 means "false" and 1, "true".

First, note that such a table has 3 degrees of freedom when the margins n_a and n_b are not fixed, that is, one can reconstruct the table from knowing only 3 values. The knowledge of 3 not linked values, for example *SUP*, *CONF* and *LIFT* completely determines the joint frequency distribution of A and B (Table 1).

Following proponents of association rules, using a support-confidence approach, attention is focused on examples A, on p_{ab} (support) and on $p_{b/a} = p_{ab}/p_a$ (confidence). Distribution of examples $\overline{A}$ between B and $\overline{B}$ is not taken into consideration.

The various examples and counter-examples of association rules, of implication and of equivalence that can be derived from A and B are displayed in Fig. 1. Rules are on level 1, implications on level 2 and equivalences on level 3. For each of the possible 8 rules, 4 implications and 2 equivalences, a $A \times B$ cross table is derived, where the values of A $(0, 1)$ are the lines and those of B $(0, 1)$ the columns. Each combination is marked as an example $(+)$, a counter-example $(-)$, or not accounted for (o). The rules, implications and equivalences on the left-hand side are positive while those on the right-hand side are negative.

The rule $A \rightarrow B$ has a single counter-example, $A\overline{B}$, and a single example, AB. One can see that a rule and its contrapositive share the same counter-examples but have different examples. The implication $A \Longrightarrow B$ and its contrapositive $\overline{B} \Longrightarrow \overline{A}$ are equivalent to $\overline{A} \vee B$, with $A\overline{B}$ as the only counter-example. Finally, the equivalence $A \Leftrightarrow B$ and its contrapositive $\overline{B} \Leftrightarrow \overline{A}$ correspond to $(AB) \vee (\overline{A}\overline{B})$; their examples (resp. counter-examples) are the examples (resp. counter-examples) of the 4 covariant rules.

2.3 A List of Measures

Table 2 lists the usual measures of interest for association rules which respect the nature of the association rules, measures that are decreasing with $n_{a\bar{b}}$, margins n_a and n_b being fixed, and distinguish $A \rightarrow B$ from $A \rightarrow \overline{B}$. In the reminder of this chapter, only those measures will be considered. Other measures are given in [23, 44, 20].

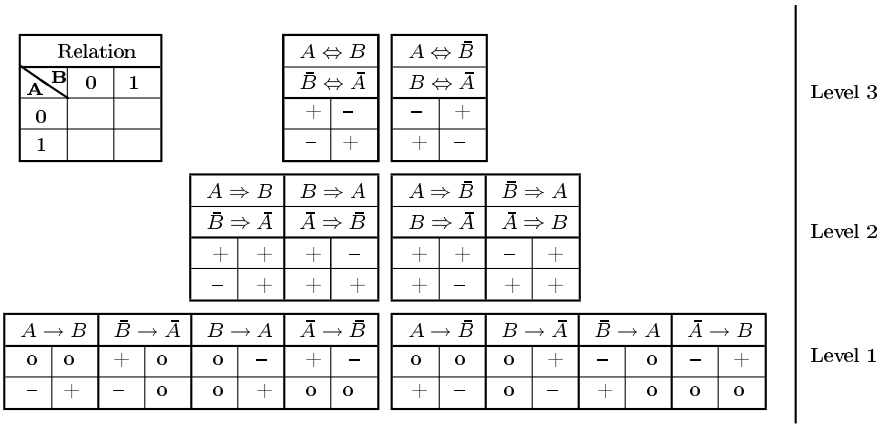

Fig. 1. Examples and counter-examples of rules, implications and equivalencies

Table 2. Usual Measures of interest

Measure	Formula	Acronym	Ref.
Support	p_{ab}	SUP	[2]
Confidence	$p_{b/a}$	$CONF$	[2]
Centered confidence	$p_{b/a} - p_b$	$CENCONF$	
Ganascia	$2p_{b/a} - 1$	GAN	[13]
Piatetsky-Shapiro	$np_a\left(p_{b/a} - p_b\right)$	PS	[39]
Loevinger	$\frac{p_{b/a} - p_b}{p_{\bar{b}}}$	LOE	[36]
Zhang	$\frac{p_{ab} - p_a p_b}{Max\left\{p_{ab}p_{\bar{b}};\, p_b p_{a\bar{b}}\right\}}$	$ZHANG$	[48]
Correlation Coefficient	$\frac{p_{ab} - p_a p_b}{\sqrt{p_a p_{\bar{a}} p_b p_{\bar{b}}}}$	R	
Implication Index	$\sqrt{n}\frac{p_{a\bar{b}} - p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$	$IMPIND$	[35]
Lift	$\frac{p_{ab}}{p_a p_b}$	$LIFT$	[11]
Least contradiction	$\frac{p_{ab} - p_{a\bar{b}}}{p_b}$	LC	[3]
Conviction	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$	$CONV$	[10]
Implication Intensity	$P\left[Poisson\left(np_a p_{\bar{b}}\right) \geq np_{a\bar{b}}\right]$	$IMPINT$	[16]
Sebag-Schoenauer	$\frac{p_{ab}}{p_{a\bar{b}}}$	SEB	[42]
Bayes Factor	$\frac{p_{ab}p_{\bar{b}}}{p_{a\bar{b}}p_b}$	BF	[25]

2.4 Assessment Criteria

A number of criteria that can be used to assess a measure will be studied, yielding a critical review of the usual measures of interest. Tan *et al.* [44] undertook a similar exercise for symmetric or symmetrized measures.

The very Meaning of a Measure

Does the measure under study have a clear, concrete meaning for the user? It is so for *SUP* and *CONF*, and also for *LIFT*, *CONV*, *SEB* or *BF*. A measure with a lift of 2 means that the number of examples of the rule $A \rightarrow B$ is twice what is expected under independence. Hence, a customer who buys A is twice as likely to buy B than the general consumer, but similarly, he who buys B is twice as likely to buy A, since lift is symmetric and that the examples of $A \rightarrow B$ are also those of $B \rightarrow A$. $CONV = 2$ means that $n_{a\bar{b}}$ is half the expected number under the independence of A and B. When $SEB = 2$, the odds of "buying B" given "A was bought" is 2, or, he who buys A is twice as likely to buy B than to not buy B, or chances of buying B are $2/3$. If $BF = 2$, odds of buying B are doubled if A is bought. Interpreting other measures is not as easy, especially *ZHANG* and *EII*, the entropic form of *IMPINT*.

Measure and Corresponding Rule

A measure must distinguish the various rules associating A and B (Fig. 1).

1. A measure must permit a clear choice between $A \rightarrow B$ and $A \rightarrow \overline{B}$, since the examples of one are the counter-examples of the other. Thus, *Pearl's*, *J-measure* and χ^2 (see [20] for those measures) were eliminated, since they do not account for the positivity or negativity of the rule.
2. Asymmetric measures which respect the nature of transactional rules are preferred: "if those items (A) are in the basket, then quite often those (B) are also". Symmetric measures like *SUP*, *PS*, *LIFT*, or R and its derivatives, give the same assessment of rules $A \rightarrow B$ and $B \rightarrow A$; while these rules have the same examples, they do not have the same counter-examples.
3. Should a measure give the same assessment to $A \rightarrow B$ and $\overline{B} \rightarrow \overline{A}$ [27]? If logical implication requires a strict equality, it is not so in the context of association rules. Indeed, both rules have the same counter-examples but not the same examples. The entropic intensity of implication, or *EII* [18] accounts for the contrapositive and brings the rule and the logical implication closer.

Examples and Counter-Examples

At first glance, one could say that a rule is unexpected whether one pays attention to the exceptionally high number of examples of the rule, n_{ab}, or

Table 3. Behaviour of certain measures in extreme situations

Situation	**Incompatibility**	**Independence**	**Logical rule**
Characterization	$p_{ab} = 0$	$p_{ab} = p_a p_b$	$p_{ab} = p_a$
Support	0	$p_a p_b$	p_a
Confidence	0	p_b	1
Centered confidence	$-p_b$	0	$p_{\overline{b}}$
Ganascia	-1	$2p_b - 1$	1
Piatetsky-Shapiro	$-np_a p_b$	0	$np_a p_{\overline{b}}$
Loevinger	$\frac{-p_b}{p_{\overline{b}}}$	0	1
Zhang	-1	0	1
Correlation Coefficient	$-\sqrt{\frac{p_a p_b}{p_{\overline{a}} p_{\overline{b}}}}$	0	$\sqrt{\frac{p_a p_{\overline{b}}}{p_{\overline{a}} p_b}}$
Implication Index (-)	$-p_b\sqrt{\frac{np_a}{p_{\overline{b}}}}$	0	$\sqrt{np_a p_{\overline{b}}}$
Lift	0	1	$\frac{1}{p_b}$
Least contradiction	$-\frac{p_a}{p_b}$	$2p_a - \frac{p_a}{p_b}$	$\frac{p_a}{p_b}$
Conviction	$p_{\overline{b}}$	1	∞
Implication Intensity	0	0.5	1
Sebag-Schoenauer	0	$\frac{p_b}{1-p_b}$	∞
Bayes Factor	0	1	∞

to the exceptionally low number of counter-examples, $n_{a\overline{b}}$. However, the examples of $A \rightarrow B$ are also those of $B \rightarrow A$ (Fig. 1), whereas the counter-examples of $A \rightarrow B$ are also those of $\overline{B} \rightarrow \overline{A}$. This justifies a preference for the counter-examples. To obtain a true difference between those options, one should, following Lerman *et al.* [35], explore the counter-examples and some probabilistic model; it is important that the margin n_a be not fixed, otherwise the number of examples and of counter-examples would be dependant, $n_{ab} + n_{a\overline{b}} = n_a$. *IMPIND*, *IMPINT* and *EII* derived from Lerman's model 3 (see Sect. 2.4) are such measures.

Direction of the Variation in the Measure and Reference Points

We limited our study to the measures that are decreasing with the number of counter-examples, margins n_a and n_b being fixed. Such a measure is maximum when $n_{a\overline{b}} = 0$, that is when $p_{b/a} = 1$, which corresponds to a logical rule. It is minimum when $n_{a\overline{b}} = n_a$, that is $n_{ab} = 0$, and $p_{b/a} = 0$, which means that A and B are incompatible. In fact, a rule is interesting whenever $n_{a\overline{b}} < \frac{n_a n_{\overline{b}}}{n}$, that is when $p_{b/a} > p_b$ ($p_{b/a} = p_b$, when A and B are independent). According to Piatetsky-Shapiro [39], a good measure should be:

a . $= 0$, A and B are independent, $p_{ab} = p_a p_b$
b . > 0, under attraction, $p_{ab} > p_a p_b$
c . < 0, under repulsion, $p_{ab} < p_a p_b$

He proposes *PS*, a symmetric measure whose bounds depend on A and B, $PS(A \rightarrow B) = np_a \left[p_{b/a} - p_b\right] = n\left[p_{ab} - p_a p_b\right]$. The conditions b and c above can be replaced by normalizing conditions b' and c' [48], which gives the so-called *ZHANG*:

b'. $= 1$, in case of a logical rule ($p_{b/a} = 1$, i.e. $A \subset B$)

c'. $= -1$, in case of incompatibility ($p_{b/a} = 0$, i.e. $AB = \emptyset$)

The only measures that take fixed reference values in the case of independence and extreme values (Table 3) are *ZHANG* and *BF*. However, the value in case of incompatibility is not very important since the only interesting situations are those where $p_b \leq p_{b/a} \leq 1$.

The lower reference point is thus often the case of independence. In that case, for the measures listed in Table 3, the value is often fixed, most often 0, sometimes 1 (*LIFT*) or 0.5 (*IMPINT*). The only exceptions are *CONF*, *LC*, *SEB* and *GAN*, or again some derived measure like the example and counter-example rate, $ECR = 1 - \frac{1}{SEB}$. As pointed out in Blanchard *et al.* [8], these are measures for which the lower reference point is not independence but rather indetermination ($n_{a\bar{b}} = \frac{n_a}{2}$, that is $p_{b/a} = p_{\bar{b}/a} = 0.5$). Lallich [28] suggested modifying *SEB* so that it be fixed under independence:

$\frac{p_{\bar{b}}}{p_b} SEB(A \rightarrow B) = \frac{p_{ab}p_{\bar{b}}}{p_{a\bar{b}}p_b} = LIFT(A \rightarrow B) \times CONV(A \rightarrow B)$.

This measure is similar to *Sufficiency* proposed by Kamber and Shingal [26]. It is actually similar to a Bayes factor [25], hence its name and notation *BF*.

On the other hand, the higher reference point is always when no counter-example exists, that is the logical rule. Normalizing to 1 is not always advisable in this case, as all logical rules are given the same interest. *LIFT* would tend to favour that of two rules which has the lower p_b.

Non-Linear Variation

Some authors [18] think it is preferable that the variation of a measure M be slow as the first counter-examples are encountered to account for random noise, then quicker, and then slow again (concave then convex). This is not the case of confidence and of all measures derived through an affine transformation which depends only of the margins n_a and n_b (Table 5). In fact, confidence is an affine function of the number of examples (or counter-examples) which depends only of n_a:

$CONF\,(A \rightarrow B) = \frac{n_{ab}}{n_a} = 1 - \frac{n_{a\bar{b}}}{n_a}$.

Conversely, to penalize false discoveries, *BF* will be preferred, as it decreases rapidly with the number of counter-examples (convex for values of $n_{a\bar{b}}$ in the neighbourhood of 0).

Impact of the Rarity of the Consequent

Following Piatetsky-Shapiro [39], a measure M must be an increasing function of $1 - p_b$ the rarity of the consequent, for fixed p_a and p_{ab}. Indeed, the rarer

the consequent B is, the more "$B \supset A$" becomes interesting. This is especially true when the support condition is not taken into consideration anymore. This is partly what happens when a measure derived from centering confidence on p_b is used. This is also obtained by merely multiplying by p_b or by dividing by $1 - p_b$; thus, the measure $BF = \frac{p_{\bar{b}}}{p_b} SEB$ improves *SEB* in this respect.

Descriptive vs. Statistical Approaches

Measures can be regarded as descriptive or as statistical [28, 19]. A measure is descriptive if it remains unchanged when all the counts are multiplied by a constant θ, $\theta > 1$. Otherwise, the measure is said to be statistical. It seems logical to prefer statistical measures, as the reliability of its assessment increases with n, the number of transactions. A statistical measure supposes a random model and some hypothesis H_0 concerning the lower reference point, quite often, the independence of A and B [35]. One can consider that the base at hand is a mere sample of a much larger population, or that the distribution of 0's and 1's is random for each item.

We denote by N_x the random variable generating n_x. Under the hypothesis of independence, Lerman *et al.* [35] suggest that a statistical measure can be obtained by standardizing an observed value, say the number of counter-examples $N_{a\bar{b}}$, giving :

$N^{CR}_{a\bar{b}} = \frac{N_{a\bar{b}} - E(N_{a\bar{b}}/H_0)}{\sqrt{Var(N_{a\bar{b}}/H_0)}}$.

This statistical measure is asymptotically standard normal under H_0. A probabilistic measure is given by $1 - X$, where X is the right tail p-value of $N^{CR}_{a\bar{b}}$ for the test of H_0, which is uniformly distributed on $[0, 1]$ under H_0.

Lerman *et al.* [35] propose that H_0 be modelled with up to 3 random distributions (Hyp, Bin, and Poi denoting respectively the hypergeometric, the binomial and the Poisson distributions):

- Mod. 1: n, n_a fixed, $N_{a\bar{b}} \equiv Hyp(n, n_a, p_{\bar{b}})$
- Mod. 2: $N_a \equiv Bin(n, p_a)$; $/N_a = n_a$, $N_{a\bar{b}} \equiv Bin(n_a, p_{\bar{b}})$
- Mod. 3: $N \equiv Poi(n)$; $/N = n$, $N_a \equiv Bin(n, p_a)$; $/N = n, N_a = n_a$, $N_{a\bar{b}} \equiv Bin(n_a, p_{\bar{b}})$

Depending on the model, $N_{a\bar{b}}$ is distributed as a $Hyp(n, n_a, p_{\bar{b}})$, as a $Bin(n, p_a p_{\bar{b}})$ or a $Poi(n p_a p_{\bar{b}})$. When standardizing, the expectation is the same, but the variance is model-dependent. Model 1 yields the correlation coefficient R, whereas Model 3 yields *IMPIND* the implication index. The latter has the advantage of being even more asymmetrical. Each statistical measure gives in turn a probabilistic measure; for example, under Model 3, $IMPINT = P(N(0,1) > IMPIND)$ [16, 17].

Discriminating Power

Statistical measures tend to lose their discriminating power when n is large as small deviations from H_0 become significant. Consider the example of Table 4

Table 4. Displaying dilatation based on one example

		(a),(b),(c)	(a),(b),(c)	(a)	(b)	(c)	(a)	(b)	(c)
p_{ab}	$p_{a\bar{b}}$	$CONF$	R	R^{cr}	R^{cr}	R^{cr}	M	M	M
0.00	0.30	0	-0.65	-2.93	-4.14	-9.26	0.002	0.000	0.000
0.05	0.25	0.17	-0.44	-1.95	-2.76	-6.17	0.025	0.003	0.000
0.10	0.20	0.33	-0.22	-0.98	-1.38	-3.09	0.165	0.084	0.001
0.15	0.15	0.5	0	0	0	0	0.500	0.500	0.500
0.20	0.10	0.67	0.22	0.98	1.38	3.09	0.835	0.916	0.999
0.25	0.05	0.83	0.44	1.95	2.76	6.17	0.975	0.997	1.000
0.30	0.00	1	0.65	2.93	4.14	9.26	0.998	1.000	1.000

where the margins are fixed, $p_a = 0.30$ and $p_b = 0.50$. Examine how the various measures react to changes in $p_{a\bar{b}}$ the proportion of counter-examples. The measures considered here are *CONF*, *R*, R^{CR} (*R* standardized under independence), and M the p-value of R under independence. The various measures are compared with $n = 20$ (columns (a)), $n = 40$ (columns (b)), and $n = 200$ (columns (c)). Clearly, as n grows, M is less able to distinguish the interesting rules. On the other hand, the ordering remains unchanged. As n is the same for all rules of a given base, one might want to first select the rules that reject independence to the benefit of positive dependence, then considered centered descriptive measures and reason on the ordering induced by those measures.

The contextual approach, developed by Lerman for classification problems, offers a first solution to the loss of discriminating power suffered by statistical measures: consider the probabilistic discriminant index *PDI* [34]. This index is defined as $PDI(A \to B) = 1 - \Phi\left[IMPINT(A \to B)^{CR/\mathcal{R}}\right]$, where Φ is the standard Gaussian distribution function and $\mathcal{R}$ is a base of admissible rules. This base can contain all the rules, or only those that meet some conditions, for example conditions on support and confidence, or even the additional condition $n_a < n_b$.

It has been suggested by Gras *et al.* [18] that the statistical measure (*IMPINT)* be weighted by some inclusion index based on the entropy H of B/A and $\overline{A}/\overline{B}$. With $H(X) = -p_x \log_2 p_x - (1-p_x)\log_2(1-p_x)$, these authors also define $H^*(X) = H(X)$, if $p_x > 0.5$, and $H^*(X) = 1$, otherwise. The inclusion index, noted $i(A \subset B)$, is then defined as

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

In later work [19, 7], the authors recommend using $\alpha = 2$ as it allows a certain tolerance with respect to the first counter-examples, and define the entropic implication index, noted *EII*, as $EII = [IMPINT \times i(A \subset B)]^{\frac{1}{2}}$.

Parameterization of Measures

As shown in [19], the lower reference situation is that of indetermination. This is preferable to independence for predictive rules. More generally, Lallich *et*

al. [31] have suggested a parameterized lower situation, adapted to the case of targeting. It can be written as $p_{b/a} = \theta$ or $p_{b/a} = \lambda p_b$. After parameterizing and rewriting the usual measures, it comes that *GAN* and *LOE* are special cases of a single parameterized measure $\frac{p_{b/a}-\theta}{1-\theta}$ for $\theta = 0.5$ and $\theta = p_b$. Moreover, each of the statistical, probabilistic and discriminant measures derived from Lerman *et al.* model-based approach [35] has been parameterized. The null hypothesis can be written as $H_0 : \pi_{b/a} = \theta$ (or possibly $\pi_{b/a} = \lambda\pi_b$), with $\pi_{b/a}$ the theoretical confidence of the rule over all possible cases, and π_b the theoretical frequency of B under a right tail alternative. In particular, under Model 3, the parameterized version of *IMPIND* and *IMPINT*, noted $IMPINDG_{|\theta}$ and $IMPINTG_{|\theta}$, are given by:

$IMPINDG_{|\theta} = \frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$;

$IMPINTG_{|\theta} = P(N(0,1) > IMPINDG_{|\theta})$.

The parameterized discriminant versions are obtained by transforming the entropy used in constructing EII into a penalizing function $\widetilde{H}(X)$, $\widetilde{H}(X) = 1$ for $p_x = \theta$ (instead of 0.5). It is sufficient, in the formula for $H(X)$, to replace p_x by $\widetilde{p}_x$, $\widetilde{p}_x = \frac{p_x}{2\theta}$, if $p_x < \theta$, and $\widetilde{p}_x = \frac{p_x+1-2\theta}{2(1-\theta)}$, if $p_x \geq \theta$. Let $\widetilde{H}^*_{|\theta}(X) = \widetilde{H}_{|\theta}(X)$, if $p_x > 0.5$, and $\widetilde{H}^*_{|\theta}(X) = 1$, otherwise. The generalized inclusion index $i_{|\theta}(A \subset B)$ is given by:

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

A generalized entropic implication index can then be derived as:

$GEII_{|\theta} = \left[IMPINT_{|\theta}(A \rightarrow B) \times i_{|\theta}(A \subset B)\right]^{\frac{1}{2}}$.

Establishing a Threshold

It is important that the measures considered allow the establishment of a threshold able to retain only the interesting rules, without resorting to classifying all of them [28]. Classically, the threshold is defined in relation to the cumulative probability of the observed measure under H_0 for a given model. Note that the threshold is not a risk level for the multitude of tests, but merely a control parameter. By definition, it is possible to set such a threshold directly for *PDI* and *IMPINT*. Other measures do not allow such direct a calculation. It is quite complex for *ZHANG* because of the standardization, and for *EII* because of the correction factor.

Ordering Induced by a Measure

Two measures M and M' give the same order to the rules of a transactional base if and only if for all pairs of rules extracted from the base:

$M(A \rightarrow B) > M(A' \rightarrow B') \Longleftrightarrow M'(A \rightarrow B) > M'(A' \rightarrow B')$

This defines an equivalence relation on the set of possible measures [28]. For example, *SEB* orders like *CONF* because it can be written as a monotonic

increasing transformation of *CONF*, $SEB = \frac{CONF}{1-CONF}$. Similarly, one can show that *LOE* orders like *CONV*, and that *PDI* orders like *IMPINT*. It should be pointed out that if the consequent is given, that is p_b fixed as it is the case for association rules in supervised learning, then *LIFT*, *CONV*, *LOE* and *SEB* order like *CONF*.

2.5 Various Measures of the Interest of a Rule

We have shown that alternatives to *SUP* and *CONF* are necessary to identify interesting rules and we have proposed several selection criteria. Now, let us specify the link between the usual measures and confidence, highlighting those that are affine transformations of confidence.

First, let's stress that the support is the index of association for Boolean variables proposed by Russel and Rao [41]; then, it will be pointed out that the usual indices of proximity defined on logical variables (see [33]) are not useful for the assessment of association rules because of their symmetrical treatment of Boolean attributes.

Affine Transformations of Confidence

Several measures can be written as a standardization of confidence via some affine transformation [28], namely $M = \theta_1 (CONF - \theta_0)$, whose parameters only depend on the relative margins of the $A \times B$ cross table and possibly on n (Table 5). Most often, the change in location indicates a departure from independence, $p_{b/a} - p_b$, and the change in scale depends on the ultimate goal. Conversely, changes in scale inform on what distinguishes two measures centered on p_b. There are two notable exceptions, *LIFT* for which the comparison to p_b is merely a change of scale and *LC* which centers confidence at 0.5.

Measures Derived from Confidence via Some Affine Transformation

All these measures improve on confidence but, by construction, inherit its principal characteristics. For fixed margins, they are affine functions of the number of counter-examples. Moreover, these measures remain invariant under changes in n when θ_1 and θ_0 parameters do not depend on n, which is the case for all of them except *PS* and *IMPIND*.

The lift is interpreted as the quotient of the observed and expected number of examples, assuming the independence of A and B. As an expression of the number of examples, it is symmetrical, since the rules $A \rightarrow B$ and $B \rightarrow A$ have the same examples (Fig. 1). *LC* is another transformation of confidence, but centered on 0.5 rather than on p_b, a better predictive that targeting tool.

Pearson's correlation R between two itemsets can be positive (see Fig. 1, examples and counter-examples of $A \Leftrightarrow B$) or negative (see Fig. 1, $A \Leftrightarrow \overline{B}$). R

Table 5. Measures derived from confidence via some affine transformation

Measure	**center (θ_0)**	**scale (θ_1)**
Centered confidence	p_b	1
Ganascia	0.5	2
Piatetsky-Shapiro	p_b	np_a
Loevinger	p_b	$\frac{1}{p_{\overline{b}}}$
Zhang	p_b	$\frac{p_a}{Max}$
Correlation Coefficient	p_b	$\frac{\sqrt{p_a}}{\sqrt{p_{\overline{a}}p_b p_{\overline{b}}}}$
Implication Index	p_b	$\sqrt{n}\sqrt{\frac{p_a}{p_{\overline{b}}}}$
Lift	0	$\frac{1}{p_b}$
Least contradiction	0.5	$2\frac{p_a}{p_b}$

is linked to the χ^2 of independence between A and B used by Brin *et al.* [10], since $\chi^2 = nR^2$, with $nR^2 \approx N(0,1)^2$, assuming independence. Contrary to χ^2, R distinguishes the cases $A \rightarrow B$ and $A \rightarrow \overline{B}$. The correlation coefficient R can be written as:

$R = \frac{p_{ab} - p_a p_b}{\sqrt{p_a p_{\overline{a}} p_b p_{\overline{b}}}} = \frac{\sqrt{p_a}}{\sqrt{p_{\overline{a}} p_b p_{\overline{b}}}} [CONF - p_b]$.

This can be simplified as $R = \frac{p_{ab} - p_b^2}{p_b p_{\overline{b}}} = \frac{1}{p_{\overline{b}}} [CONF - p_b] = LOE$, when A and B have the same marginal distribution ($p_a = p_b$), and as $R = 2CONF - 1$ (i.e. *GAN*), when this distribution is balanced ($p_a = p_b = 0.5$). Thus, R and *CONF* can be seen as equivalent for cross tables with balanced margins. The cross table is then symmetrical, meaning that the 4 covariant rules connecting A and B or their complement have the same confidence, as well as the 4 contravariant rules.

Other Measures

The measures that cannot be reduced to an affine transformation of confidence are *CONV*, *SEB* and *BF*, as well as measures derived from the implication index. CONV can be expressed as a monotonic increasing function of *LOE*, $CONV = (1 - LOE)^{-1}$. *CONV* is analogous to *LIFT* applied to the counter-examples:

$CONV(A \rightarrow B) = \frac{p_{\overline{b}}}{p_{\overline{b}/a}} = \frac{p_a p_{\overline{b}}}{p_{a\overline{b}}} = LIFT(A \rightarrow \overline{B})^{-1}$.

SEB is a monotonic increasing transformation of confidence, as well as – just like *BF* – an affine transformation of conviction with fixed margins:

$$SEB = \frac{CONF}{1 - CONF} = \frac{1}{p_{\overline{b}}} (CONV - 1); \; BF = \frac{1}{p_b} (CONV - 1) \; .$$

Statistical measures are based on *IMPIND* which is an affine transformation of *CONF* with fixed margins, namely $IMPINT = P(N(0,1) > IMPIND)$, and its discriminating versions *EII* and *PDI*.

Strategy

The user must choose the measures the most appropriate to his objective and to the characteristics of his data; criteria proposed in this section may be found of help. The user can also opt for some automated decision-making procedure to decide on the most appropriate measure [32].

Because support and confidence are more easily understood, and because support condition is antimonotonic, support-confidence algorithms are often applied first to transactional databases. A set of admissible rules for the selected support and confidence thresholds is then obtained. Such sets comprise a large number (m) of rules, not always interesting. The most interesting rules can be identified with the help of the selected measures. If the support condition is released and if the interesting rules are sought for directly with the selected measures, the number of rules becomes excessively large. Then, one may be restricted to simple rules [3].

3 Validating Interesting Rules

Rules that are truly interesting for the user are those for which the real world value, for the selected measure, exceeds some preset threshold. Most often, as the transactional database is seen as a mere sample of the universe of all possible transactions, one only knows some empirical evaluation of those rules. The problem becomes the selection of the rules whose empirical values significantly exceed the threshold. This means testing each one of the m rules, that is, m tests.

For example, one can seek rules significantly far from the independence of A and B, which leads to selecting rules for which confidence $p_{b/a}$ is significantly larger than the threshold p_b. The hypothesis of independence, noted H_0, is given by $\pi_{b/a} = \pi_b$, where $\pi_{b/a}$ is the theoretical confidence (or confidence over all possible transactions), whereas π_b is the prior theoretical frequency of B. For each rule, H_0 is tested against the right-tail alternative of positive dependence noted H_1 given by $\pi_{b/a} > \pi_b$. If one is seeking predictive rules, one would select rules for which the confidence $p_{b/a}$ is significantly larger than 0.5, that is, testing $H_0 : \pi_{b/a} \leq 0.5$ against $H_1 : \pi_{b/a} > 0.5$. If the objective is targeting of group, one can also seek rules for which the confidence $p_{b/a}$ is significantly larger than the threshold λp_b, that is, a lift larger than some set value $\lambda > 1$; this is equivalent to testing $H_0 : \pi_{b/a} \leq \lambda \pi_b$ against $H_1 : \pi_{b/a} > \lambda \pi_b$.

These various situations could be analyzed with a measure other than confidence. If q measures are available, one needs a total of qm tests. Moreover, certain measures have a complicated algebraic expression (e.g. *EII*) which impedes the elaboration of a parametric test. In summary, the validation of interesting rules requires the ability to develop a multitude of tests using some possibly non-parametric device.

This multiplicity of tests inflates the number of false discoveries (rules wrongly selected). Indeed, if m tests are developed, each with a probability of Type I error set at α_0, even if no rule is truly interesting, the procedure creates on average $m\alpha_0$ false positives. Controlling multiple risk is rarely a topic in data mining literature. A noteworthy exception is the work of Meggido and Srikant [38] on the significance of association rules with respect to independence, who simulate the number of false discoveries for a given level of Type I risk. On the other hand, this topic is well covered in biostatistics (see Sect. 3.1). The authors have proposed in earlier work methods to control multiple risk using statistical learning theory and VC-dimension [45], or bootstrap [29]. In practice, because they make no allowance for false discoveries among the m rules, these methods have little power, yet ignoring significant rules. The authors have proposed *BS_FD* [30] to test the significance of rules; this method controls the number of false discoveries and uses an original bootstrap criterion. The general case with any threshold is exposed below.

First, the problem of controlling risk with multiple tests will be reviewed (Sect. 3.1), as well as procedures that control risk using p-values (Sect 3.2). Then, *BS_FD* will be introduced (Sect. 3.3) and will be applied to selecting the most interesting association rules (Sect. 3.4).

3.1 Constructing Multiple Tests

Significance Test for a Rule

Consider a rule $A \rightarrow B$ and some measure of interest M, decreasing with $n_{a\bar{b}}$ and fixed margins. Note M_{obs} the observed value of $M(A \rightarrow B)$ on the sample of transactions and μ its theoretical value on a very large set of transactions. The rule is said to be significant under M with respect to μ_0 if $M_{obs} = M(A \rightarrow B)$ is significantly larger to some preset value μ_0. A test for the null hypothesis $H_0 : \mu = \mu_0$ against the unilateral alternative $H_1 : \mu > \mu_0$ is needed. H_0 is rejected whenever M_{obs} is too far from H_0 in the direction of H_1, with a Type I error risk set at $\alpha = \alpha_0$. The p-value for M_{obs} is computed as the probability of obtaining a value as large as M_{obs} assuming H_0 is true, and the rule is selected if the p-value for M_{obs} is less than α_0. Obviously, this requires the knowledge of the distribution of $M(A \rightarrow B)$ under H_0.

Risk and Type I Error

The identification of the significant rules under M among the m rules extracted from a transactional database requires m tests. This raises the problem of false discoveries, a recurrent problem in data mining. If m uninteresting rules are tested at the level α_0, then, on average, $m\alpha_0$ rules will mechanically be erroneously selected. For example, with $\alpha_0 = 0.05$, and a base of extracted rules comprising $m = 10,000$ rules, even if all were non-significant, about 500 rules would mechanically be selected.

Table 6. Synthesis of the results of m tests

Reality \ Decision	Acceptation	Reject	Total
H_0 true	U	V	m_0
H_1 true	T	S	m_1
Total	W	R	m

To take into account the multiplicity of tests, the fundamental idea of Benjamini and Hochberg [4] is to consider the number of errors over m iterations of the test, rather than the risk of being wrong on one test (see Table 6, where a upper case represents observable random variates and lower case are fixed yet unknown quantities m_0 and m_1). From this table, these authors derive several indicators. Two most common ones are described next, *FWER* (Family wise error rate) and *FDR* (False Discovery Rate).

FWER is the chance of erroneously rejecting H_0 at least once, $FWER = P(V > 0)$. It is a much too strict criterion for a large number of tests, because it does not allow any false discovery.

The authors [30] proposed the *User Adjusted Family Wise Error Rate*, $UAFWER = P(V > V_0)$, an original and more flexible variant which allows V_0 false discoveries. $UAFWER$ can be controlled at the level δ using a bootstrap-based algorithm (Sect. 3.3).

Several quantities using the expectation of V, the number of false discoveries, possibly standardized, have been proposed to remedy the difficulties inherent to *FWER*. The best known is *FDR* [4], the expected proportion of erroneous selections among the selected rules. When $R = 0$, define $\frac{V}{R} = 0$, that is, $FDR = E(Q)$, where $Q = \frac{V}{R}$ if $R > 0$, 0 otherwise. Then:

$FDR = E(\frac{V}{R} \mid R > 0)P(R > 0)$.

Storey [43] proposed the *pFDR*, a variation of *FDR*, using the knowledge that H_0 has been rejected at least once:

$pFDR = E(\frac{V}{R} \mid R > 0)$.

At the cost of a fixed proportion of erroneous selections, these quantities are less severe, thus augmenting the probability of selecting an interesting rule (increased power). One has $FDR \leq FWER$ and $FDR \leq pFDR$, hence $FDR \leq pFDR \leq FWER$ when m is large, because $P(R > 0)$ goes to 1 as m increases. The problem of controlling the Type I risk is resolved in the literature by the use of p-values. *FWER* and *FDR* will be examined in turn.

3.2 Controlling Multiple Risk with p-values

Several solutions have been proposed to control *FWER* or *FDR*, most recently in the context of gene selection. A remarkable summary of this work can be found in [14].

Control of FWER

Bonferroni Correction

Let us denote by P_r the random variable generating the p-value p_r associated to the test statistics $T_r, r = 1, \ldots, m$. One can show that $FWER = 1 - P(\bigcap_{r=1}^{m} (P_r > \frac{\alpha_0}{m}) \mid H_0)$. Assuming that the rules are independent, then $FWER = 1 - (1 - \frac{\alpha_0}{m})^m \approx \alpha_0$. The Bonferroni correction consists in constructing each test at the level $\frac{\alpha_0}{m}$, in order to set the $FWER$ on α_0. This correction is usually applied by adjusting the m p-values. The adjusted p-value $\widetilde{p}_r$ is defined by $\widetilde{p}_r = \min\{mp_r, 1\}$. All rules having an adjusted p-value smaller than the risk α_0 are selected. If independence cannot be assumed, one only has $\frac{\alpha_0}{m} \leq FWER \leq \alpha_0$. The Bonferroni correction is not a good solution for two reasons:

- FWER is actually not controlled, but somewhere between $\frac{\alpha_0}{m}$ and α_0; it is equal to α_0 only when the rules are mutually independent. Now, rules are not independent, as they share items and because items are dependent.
- FWER is conservative, thus increasing the risk of a Type II error, that is not finding an interesting rule.

Holm's Step-down Procedure

Stepwise procedures examine p-values in increasing order of magnitude, adjusting the critical value as the procedure progresses. Holm [24] considers that a selected variable corresponds to H_0 false, and the critical value is adjusted to only account for the variables remaining to be examined. Since the p-values are sorted in increasing order, with $p_{(i)}$ the i^{th} p-value, H_0 is rejected while $p_{(i)} < \frac{\alpha_0}{m-i+1}$. H_0 is accepted for all p-values following the first acceptance. This procedure, easy to implement, gives good results when the number of tests is rather small, as the adjustment to the critical value has some importance. The procedure is ill-adapted to large numbers of tests.

Control of FDR

Benjamini and Liu's Procedure

Benjamini and Liu [5] proposed a sequential method for the control of FDR under the assumption of independence. The p-values are examined in increasing order and the null hypothesis is rejected if the p-value at hand $p_{(i)}$ is less than $\frac{i\alpha_0}{m}$. This procedure ensures that $FDR = \frac{m_0}{m}\alpha_0$ under independence. It is compatible with positively dependent data.

pFDR

In order to estimate $pFDR = E(\frac{V}{R} \mid R > 0)$, the proportion of false detections, Storey [43] proposes the approximation:

$\widehat{pFDR}(\delta) = \frac{\hat{\pi}_0.m.\delta}{\#\{p_i \leq \delta, i=1,\ldots,m\}}$, where

- m is the number of rules to be tested; δ defines the rejection area: hypotheses corresponding to p-values less than or equal to δ are rejected;
- p_i is the i^{th} largest p-value;
- $\pi_0 = \frac{m_0}{m}$ is the proportion of null hypotheses; here, π_0 is estimated by $\hat{f}(1)$, where $\hat{f}$ is a cubic spline of $\hat{\pi_0}(\lambda)$ over λ: $\hat{\pi_0}(\lambda) = \frac{\#\{p_i \geq \lambda, i=1,\ldots,m\}}{m(1-\lambda)}$; $0 < \lambda < 0.95$ represents the acceptation area.

The *pFDR* is defined in terms of a preset rejection area. Once the global *pFDR* is computed, variables are controlled by a step-down procedure using the q-values defined for each p-value as $\hat{q}(p_m) = \hat{\pi_0}.p_m$ and:

$\hat{q}(p_i) = \min\left(\frac{\hat{\pi_0}.m.p_i}{i}, \hat{q}(p_{i+1})\right); i = m-1, \ldots, 1$.

The q-value is to the *pFDR* what the p-value is to Type I error, or what the adjusted p-value is to the FWER. Any rule whose p-value has a corresponding q-value less than *pFDR* is selected.

3.3 Controling UAFWER Using the *BS_FD* Algorithm

We have proposed a bootstrap-based non-parametric method to control *UAFWER*. This method does not require p-values, which is advantageous when the distribution of $M(A \rightarrow B)$ under H_0 is unknown (e.g. the discriminant versions of the statistical measures, like *EII* [18] or its generalization *GEII* [31]).

Notations

- $\mathcal{T}$: set of transactions, $n = Card(\mathcal{T})$, p: number of items;
- $\mathcal{R}$: base of admissible association rules with respect to some predefined measures, for example, support and confidence, $m = Card(\mathcal{R})$;
- M: measure of interest; $\mu(r)$: theoretical value of M for rule r; $M(r)$: empirical value of M for r on $\mathcal{T}$;
- V: number of false discoveries, δ: risk level of the control procedure, with V_0 the number of false discoveries not to be exceeded given δ, $\mathcal{R}^*$ a subset of $\mathcal{R}$ comprising the significant rules as determined by M and μ_0.

Objective

The objective is to select the rules r of $\mathcal{R}$ that are statistically significant as measured by M, meaning that $M(r)$ is significantly larger than $\mu_0(r)$, the expected value under H_0. We have suggested various algorithms that use the tools of statistical learning so that 100% of the identified rules be significant for a given α, among others the bootstrap-based algorithm *BS* [29]. Experience has shown that this approach might be too prudent, therefore not powerful enough. Allowing a small number of false discoveries, after Benjamini's work (Sect. 3.1), the authors propose *BS_FD*, an adaptation of *BS* that controls the number of false discoveries.

BS_FD selects rules so that $UAFWER = P(V > V_0)$, which ensures that the number of false discoveries does not exceed V_0 at the level δ. The algorithm guarantees that $P(V > V_0)$ converges to δ when the size of the samples of transactions increases.

Algorithm *BS_FD*

Given $\mathcal{T}$, $\mathcal{R}$, and M, $\mu(r) > \mu_0(r)$ is guaranteed by setting $\mu(r) > 0$, without loss of generality simply by shifting $\mu(r)$ to $\mu(r) - \mu_0(r)$. V_0 false discoveries are allowed at risk δ. Finally, $\#E = Card(E)$.

1. *Empirical assessment.* All rules of $\mathcal{R}$ are measured using M on the set of transactions $\mathcal{T}$, creating the $M(r), r \in \mathcal{R}$.
2. *Bootstrap.* The following operations are repeated l times:
   - Sample with replacement and equal probability m transactions from $\mathcal{T}$, thus creating $\mathcal{T}'$, $Card(\mathcal{T}') = Card(\mathcal{T})$. Some transactions of $\mathcal{T}$ will not be in $\mathcal{T}'$ while some others will be there many times. All rules are measured using M, creating the $M(r)$, $r \in \mathcal{R}$.
   - Compute the differences $M'(r) - M(r)$, then compute $\varepsilon(V_0, i)$, the smallest value such that $\#\{M'(r) > M(r) + \varepsilon(V_0, i)\} \leq V_0$. Hence, $\varepsilon(V_0, i)$ is the $(V_0 + 1)^{\text{st}}$ largest element of the $M'(r) - M(r)$, during the i^{th} iteration, $i = 1, 2...l$.
3. *Summary of bootstrap samples.* There are l values $\varepsilon(V_0, i)$. Compute $\varepsilon(\delta)$, $(1-\delta)^{\text{th}}$ quantile of the $\varepsilon(V_0, i)$: that is, $\varepsilon(V_0, i)$ was larger than $\varepsilon(\delta)$ only $l\delta$ times in l.
4. *Decision.* Keep in $\mathcal{R}^*$ all rules r such that $M(r) > \varepsilon(\delta)$.

Rationale

Bootstrap methods [12] approximate the distance between the empirical and true distributions by the distance between the bootstrap and empirical distributions. At the i^{th} bootstrap iteration, there are V_0 rules whose evaluation augments by more than $\varepsilon(V_0, i)$. Given the definition of $\varepsilon(\delta)$, the number of rules whose evaluation augments by more than $\varepsilon(\delta)$ is larger than V_0 in a proportion δ of the l iterations. Consequently, selecting rules for which $M(r)$ exceeds $\varepsilon(\delta)$, one is guaranteed to have at most V_0 false discoveries at the risk level δ.

Moreover, bootstrap-based methods have solid mathematical foundations [15] which require a clearly posed question. Formally, the objective is that the distribution function of the number of rules such that $\mu(r) < 0$ while $M(r) > \epsilon$, be at least $1 - \delta$ for V_0. One gets $\#\{\mu(r) \leq 0 \text{ et } M(r) > \epsilon\} \leq \#\{M(r) \geq \mu(r) + \epsilon\}$. Theorems on bootstrap applied to a family of functions verifying the minimal conditions [47] yield the approximation of this quantity by $\#\{M'(r) \geq M(r) + \epsilon\}$, which serves as a basis for $\varepsilon(V_0, i)$ and $\varepsilon(\delta)$ described in this section.

Extension to Multiple Measures

In practice, more than one measure will be of interest, for example, *SUP*, *CONF* and a measure of the departure from independence. The extension of *BS_FD*, noted *BS_FD_mm*, is achieved by using as a summary measure the minimum of the various measures. Hence, for 3 measures M_1, M_2 and M_3, one considers $M(r) = min\{M_1(r), M_2(r), M_3(r)\}$. Using *BS_FD_mm* on M at the level δ will select rules which comply with M_1, M_2 and M_3, at level δ.

Risk of Type II errors can be optimized by working with Hadamard differentiable transformations of the M_i that will make the measures homogenous [47], for example, p-values or reductions, through standardization.

Complexity of *BS_FD*

The complexity of *BS_FD* is proportional to $l \times m \times n$, assuming that the random number generator operates in constant time. In fact, the complexity of the search for the k^{th} largest element of a table is proportional to the size of the table. The value of l must be large enough so that the finiteness of l impede not the global reliability, and be independent of both m and n. The algorithm is globally linear in $m \times n$, to a constant l linked to the bootstrap.

3.4 Application to the Rules Selection and Experimentation

Selecting Significant Rules According to Independence

Description of Data

The filtering methods presented here were applied to five sets of rules available on HERBS [46]. They were extracted using Borgelt and Kruse's implementation [9] of *Apriori* applied on data sets available from the UCI site [6]: Contraceptive Method Choice (*CMC*), Flags (*Flags*), Wisconsin Breast Cancer (*WBC*), Solar Flare I (*SFI*) and Solar Flare II (*SFII*). The authors computed for each method, the reduction rate of each set of rules after removal of non-significant rules.

Parameterization

For the "5% control", Holm and Bonferroni procedures (cf. Sect. 3.2) were applied with a level of 5%. $pFDR$ is calculated with a rejection rate of 0.1%. The number of false discoveries is shown between parentheses. The rejection zone is chosen so that it will be as acceptable as possible. FDR, described in Sect. 3.2 is used with a threshold set by the last q-value selected by $pFDR$, shown in brackets. Indeed, to compare $pFDR$ and FDR, control levels should be close. Control level for *BS_FD(R)* is set at 5%, with V_0 equal to the result of $pFDR$, on the correlation coefficient R tested against 0.

Table 7. Filtering of some sets of rules

Characteristics	CMC	Flags	WBC	SF I	SF II
# cases	1,473	194	699	323	1066
# rules	2,878	3,329	3,095	5,402	3,596
Covering rate	100%	100%	96.2%	100%	100%
Re-covering rate	259	1,848	646	1,828.6	2,277
Support threshold	5%	50%	10%	20%	20%
Confidence threshold	60%	90%	70%	85%	85%
Results	CMC	Flags	WBC	SF I	SF II
level 5%	1,401	2,181	3,094	2,544	2,558
pFDR	916 (3)	1,200 (3)	3,095 (0)	900 (5)	1,625 (4)
FDR (Benj.)	913 (0.003)	1,198 (0.0027)	/	899 (0.006)	1,626 (0.0022)
BS_FD(R)	794	1,074	3,093	604	738
Holm	742	564	3,094	432	1,020
Bonferroni	731	539	3,042	427	1,006

Results

Table 7 requires some explanations. No filter is efficient on *WBC*. This is because only one rule of the starting set has a p-value above 0.05 (viz. 0.184), an other one is at 0.023, the remaining p-values are less than 0.01, and 3,036 of them are less than 0.00005 !

For the 4 other set of rules, merely repeating independence tests shrinks the sets by 51%, 34%, 53%, and 39%. However, not all false discoveries are eliminated. Using control on multiple risk reduces the number of rules selected.

Among those, Bonferroni correction is the most stringent. It produces reductions of 75%, 81%, 65% and 60%. Though stringent, it lacks power, avoiding false positives but creating false negatives. Holm's procedure gives similar results; it is inefficient because of the large number of rules which renders the step-wise correction inoperative.

On the other hand, *pFDR*, *FDR* and *BS_FD* give moderately better results, what was expected. *BS_FD* appears to be the most stringent of the 3, especially on *Solar Flare II*. The reason is that the parameterization of *pFDR* and *FDR* ensures an average number of false discoveries equal to V_0, whereas *BS_FD* ensures that V_0 be exceeded only 0.05 of the time, which is quiet demanding. These three methods are efficient rule filters. *BS_FD* is the most complex, but is advantageously non-parametric (see next section for an example).

Thus, a filtering procedure based on controlling multiple risk eliminates that would otherwise be selected by a variety of measures. Logical rules whose consequent is very frequent (e.g. *Solar Flare II*) is an example of such measures. These attain a maximum under any measure that give a fixed maximum value to logical rules, though they present little interest and their p-values are non-significant. Conversely, computing p-values is independent of any sub-

Table 8. Predictive rules selected on CMC by applying BS_FD on LIFT and $GEII(2p_b)$

		LIFT		
		Selected	Not Selected	
$GEII_{\|2p_b}$	Selected	8	19	27
	Not Selected	6	11,042	11,048
		14	11,061	11,075

sequent ranking of the rules by descriptive measures that favour the more interesting rules, for example, asymmetric measures that favour rules with low frequency consequent.

Selecting Targeting Rules

Contrary to methods like FDR or $pFDR$, *BS_FD* does not require prior knowledge of the distribution of the measure under the null hypothesis. *BS_FD* can thus be applied to algebraically complex measures.

To illustrate this, let's turn our attention to targeting rules. These are rules for which knowing the antecedent multiplies by some constant λ the probability of observing the consequent. Here, we use $\lambda = 2$, which amounts to testing $H_0 : \pi_{b/a} = 2\pi_b$. To assess this type of rule, 2 measures are used, *LIFT* and $GEII(2p_b)$ (generalized entropic intensity index with parameter $2p_b$); the null distribution of *GEII* is not known. Under H_0, these measures are respectively 2 and 0.

The CMC base [6] was used for this experiment. First, using *Tanagra* [40] implementation of *Apriori*, 13,864 rules with a support exceeding 0.05 were extracted. Among those, the 2,789 rules for which $p_b > 0.5$ were removed. Of the 11,075 remaining rules, *BS_FD* was applied on *LIFT* and on $GEII(2p_b)$ by comparing the results to 2 and 0 respectively. Results are displayed in Table 8.

LIFT and $GEII(2p_b)$ select respectively 14 and 27 rules of the 11,075 extracted by *Apriori*. These rules ensure that B has twice as many chances of occurring if A is realized. The small number of rules allows expert examination. These are especially interesting in marketing and health sciences. In this latter case, the consequent is the occurrence of a disease, and the antecedents are possible factors of this disease. The proposed procedure detects factors that multiply notably the risk of disease.

Moreover, results show that *LIFT* and $GEII(2p_b)$ do not select the same rules (only 8 are common). *BS_FD* applied to *LIFT* naturally selects the rules with the highest measures. Thus, of the 224 rules with a *LIFT* over 2, it retains those with a value above 2.479. Among those, *BS_FD* applied to $GEII(2p_b)$ does not select those rules with $p_a > 0.2$ and $p_b < 0.1$. Contrarily, it selects those with a low frequency antecedent. Using many measures allows different assessment of the interest of a given rule.

4 Conclusion and Perspectives

Means to search for association rules in databases is one of the principal contributions of data mining compared to traditional statistics. However, the usual extraction algorithms yields a very large number of not-all-interesting rules. On the other hand, these rules overfit the data [38], which makes them hard to generalize. This double problem calls for a double solution: a careful choice of the measure of interest and retaining those rules that are significant for the objective at hand. The authors have suggested a number of criteria to help the user to choose the most appropriate measure. To avoid overfitting, the significance of each rule must be tested raising the problem of controlling multiple risk and avoiding false discoveries. To this end, the authors suggest a bootstrap-based method, *BS_FD*; the proposed method controls the risk of exceeding a fixed number of false discoveries, accounting for the dependency among the rules, and allowing the test of several measures at once. *BS_FD* can be used for filtering rules where the antecedent increases the probability of the consequent (positive dependence), for filtering targeting rules, or filtering predictive rules. Experiments show the effectiveness and efficiency of the proposed strategy. An extension of this work to filtering discriminant rules in the context of genomics is being planned.

Acknowledgement. The authors gratefully acknowledge the referees for their helpful comments and Jean Dumais, Methodology Branch, Statistics Canada, for his help with the translation of this paper.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th Very Large Data Bases Conference*, pages 487–499. Morgan Kaufmann, 1994.
2. R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *ACM SIGMOD Int. Conf. on Management of Data*, pages 207–216, 1993.
3. J. Azé and Y. Kodratoff. A study of the effect of noisy data in rule extraction systems. In *Sixteenth European Meeting on Cybernetics and Systems Research*, volume 2, pages 781–786, 2002.
4. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B*, 57:289–300, 1995.
5. Y. Benjamini and W. Liu. A step-down multiple-hypothesis procedure that controls the false discovery rate under independance. *J. Stat. Planng Inf.*, 82: 163–170, 1999.
6. C.L. Blake and C.J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.
7. J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. Mesure de la qualité des règles d'association par l'intensité d'implication entropique. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1):33–45, 2004.

8. J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium. In *ASMDA'05*, pages 191–200, 2005.
9. C. Borgelt and R. Kruse. Induction of association rules: APRIORI implementation. In *15th Conf. on Computational Statistics*, 2002.
10. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In *ACM SIGMOD/PODS'97*, pages 265–276, 1997.
11. S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, *ACM SIGMOD 1997 Int. Conf. on Management of Data*, pages 255–264, 1997.
12. B. Efron. Bootstrap methods: Another look at the jacknife. *Annals of statistics*, 7:1–26, 1979.
13. J.-G. Ganascia. Deriving the learning bias from rule properties. *Machine intelligence*, 12:151–167, 1991.
14. Y. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analysis. Tech. rep. 663, Univ. of California, Berkeley, 2003.
15. E. Giné and J. Zinn. Bootstrapping general empirical measures. *Annals of probability*, 18:851–869, 1984.
16. R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* PhD thesis, Université de Rennes I, 1979.
17. R. Gras and A. Lahrer. L'implication statistique : une nouvelle méthode d'analyse des données. *Math. Inf. et Sc. Hum.*, 120:5–31, 1993.
18. R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *EGC 2001*, 1(1-2):69–80, 2001.
19. R. Gras, R. Couturier, M. Bernadet, J. Blanchard, H. Briand, F. Guillet, P. Kuntz, R. Lehn, and P. Peter. Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1), 2004.
20. F. Guillet. Mesure de la qualité des connaissances en ECD. Tutoriel 4e Conference Extraction et Gestion des Connaissances, EGC'04, 2004.
21. P. Hajek and J. Rauch. Logics and statistics for association rules and beyond. Tutorial PKDD'99, 1999.
22. P. Hajek, I. Havel, and M. Chytil. The GUHA method of automatic hypotheses determination. *Computing*, (1):293–308, 1966.
23. R.J. Hilderman and H.J. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. *LNCS*, 2035:247–259, 2001.
24. S. Holm. A simple sequentially rejective multiple test procedure. *J. Statistic.*, 6:65–70, 1979.
25. H. Jeffreys. Some test of significance treated by theory of probability. In *Proc. Of the Cambridge Phil. Soc.*, pages 203–222, 1935.
26. M. Kamber and R. Shingal. Evaluating the interestingness of characteristic rules. In *Proceedings of KDD'96*, pages 263–266, 1996.
27. Yves Kodratoff. Quelques contraintes symboliques sur le numérique en ECD et en ECT. Lecture Notes in Computer Science, 2000.
28. S. Lallich. Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches – Université Lyon 2, 2002.

29. S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des règles d'association. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1): 193–217, 2004.
30. S. Lallich, E. Prudhomme, and O. Teytaud. Contrôle du risque multiple en sélection de règles d'association significatives. *RNTI-E-2*, 2:305–316, 2004.
31. S. Lallich, B. Vaillant, and P. Lenca. Parametrised measures for the evaluation of association rule interestingness. In *Proc. of ASMDA'05*, pages 220–229, 2005.
32. P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Évaluation et analyse multicritère des mesures de qualité des règles d'association. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1):219–246, 2004.
33. I. Lerman. Comparing partitions, mathematical and statistical aspects. *Classification and Related Methods of Data Analysis*, pages 121–131, 1988.
34. I.C. Lerman and J. Azé. Indice probabiliste discriminant de vraisemblance du lien pour des données volumineuses. *Mesures de Qualité pour la Fouille de Données*, (RNTI-E-1):69–94, 2004.
35. I.C. Lerman, R. Gras, and H. Rostam. Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75):5–35, 5–47, 1981.
36. J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4), 1947.
37. H. Mannila, H. Toivonen, and A.I. Verkamo. Efficient algorithms for discovering association rules. In *Proc. AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)*, pages 181–192, 1994.
38. N. Meggido and R. Srikant. Discovering predictive association rules. *Knowledge Discovery and Data Mining (KDD-98)*, pages 274–278, 1998.
39. G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.
40. R. Rakotomalala. Tanagra. http://eric.univ-lyon2.fr/∼ricco/tanagra, 2003.
41. P. F. Russel and T. R. Rao. On habitat and association of species of anopheline larvae in south-eastern Madras. *J. Malar. Inst. India*, (3):153–178, 1940.
42. M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In J. Boose, B. Gaines, and M. Linster, editors, *Proc. of EKAW'88*, pages 28–1 – 28–20. 1988.
43. J. D. Storey. A direct approach to false discovery rates. *J. R. Statisc. Soc., Series B*, 64:479–498, 2002.
44. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Eighth ACM SIGKDD Int. Conf. on KDD*, pages 32–41, 2002.
45. O. Teytaud and S. Lallich. Bornes uniformes en extraction de règles d'association. In *CAp'01*, pages 133–148, 2001.
46. B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In R. Bisdorff, editor, *HCP'2003*, pages 187–191, 2003.
47. A. Van Der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer Series in Statistics. Springer-Verlag Publishers, 1996.
48. T. Zhang. Association rules. In *Proceedings of PAKDD 2000*, volume 1805 of *LNCS*, pages 245–256. Springer, 2000.