Statistical Strategies for Pruning All the Uninteresting Association Rules [1]

Gemma C. Garriga [2]

Abstract. We propose a general framework to formalize the problem of capturing the intensity of implication for association rules through statistical metrics. In this framework we present properties that influence the interestingness of a rule, analyze the conditions that lead a measure to perform a perfect prune at a time, and define a final proper order to sort the surviving rules. We will discuss why none of the currently employed measures can capture objective interestingness, and just the combination of some of them in a multi-step fashion, can be reliable. In contrast, we propose a new simple modification of the Pearson coefficient that will meet all the necessary requirements. We statistically infer the convenient cut-off threshold for this new metric by empirically describing its distribution function through simulation. Experiments show a promising behaviour of our proposal.

1 PROBLEM FORMULATION

One of the most relevant tasks in Knowledge Discovery in Databases is mining for association rules in large masses of data, as it was first formulated by [1]. This task is often decomposed into two separate phases: 1/ Finding all the frequent itemsets having support over a user-specified threshold, and, 2/ Generating the association rules from the maximal discovered frequent itemsets.

The input of a frequent sets algorithm is a database $\mathcal{D}$, composed of a collection of *transactions*, where each transaction is a subset of a given fixed set of items $\mathcal{I} = \{i_1, i_2, \ldots, i_N\}$. Let $I \subset \mathcal{I}$ be an itemset, and let $Pr(I, \mathcal{D})$ be the ratio of the number of transactions in which I appears to the number of all transactions in $\mathcal{D}$, i.e. $Pr(I, \mathcal{D}) = \frac{trans(I,\mathcal{D})}{|\mathcal{D}|}$. We note the *support* of an itemset I as $Pr(I, \mathcal{D})$. An itemset is called *frequent* if its support exceeds a given user-specified threshold, σ.

In the second phase, association rules are constructed from those maximal frequent sets. In brief, given any maximal frequent itemset Z, an association rule is an expression $X \Rightarrow Y$ where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, $X \cap Y = \emptyset$ and $X \cup Y = Z$. The number of these extracted implications is usually very large, leading to a rule quality problem: just a small portion of them are interesting and the rest may be misleading. Currently, this problem can be faced by calculating an interestingness measure over the rules with the aim of statistically determining their quality. This is a common technique used by many authors ([3, 4, 5, 10, 11, 13] ...), opposed to other deterministic techniques such as grouping together related rules ([7]), or using closed itemsets to generate a non-redundant set of rules ([2]).

An order induced by a measure in a given database $\mathcal{D}$ is a total order, and in current applications the user specifies a threshold to split the sorted rules in two classes: those rules ranking under the user-specified threshold are considered uninteresting and will be pruned; the rest of rules will be considered interesting. This is a risky step since the measure might be unreliable in capturing the quality of the rule and so, some uninteresting rules can still hold while other interesting ones could be eliminated.

For the study of association rules, we also consider an asymmetric framework where one variable causes another. So, there is a need to distinguish the strength of implication of the rule $r = X \Rightarrow Y$ from its reversed $\hat{r} = Y \Rightarrow X$. The calculation and interpretation of asymmetric measures depend on which variable is considered dependent, or in other words, which part of the original itemset will be the best consequent of the rule. These kind of measures that assign different values to the two rules $X \Rightarrow Y$ and $Y \Rightarrow X$ will be called *symmetry breaking*.

2 GENERAL FRAMEWORK FOR PRUNING ASSOCIATION RULES

In this section we try to do a generalization of all the different properties and considerations stated in the broad current literature ([5, 9, 10, 11, 13] among others). Given any interestingness measure IM, we consider two objective properties making IM an accurate metric in the assessment of association rules:

P1: IM must test independence of a rule r
P2: IM must distinguish the strength of implication of a rule r against its reversed $\hat{r}$

The first property P1 derives from a common principle in association rule mining: the greater the support, the better the itemset. As authors in [4] argue, this fact is true to some extent because itemsets with high support are a source of misleading rules: they appear in most of the transactions and any other itemset (despite the meaning) seems to be a good predictor of the presence of the high-support itemset. As a consequence, most of these rules turn to be useless despite having high support and accuracy, because they hold with negative dependence or independence between antecedent and consequent.

So, property P1 states that any accuracy measure must test independence between antecedent and consequent of a rule. Formally, this means that $IM(A \Rightarrow B) = k$ when $Pr(A \cup C, \mathcal{D}) = Pr(A, \mathcal{D}) \times Pr(C, \mathcal{D})$ (where k can be any constant value), and it was first formulated by [11]. So, we want that IM can clearly distinguish rules according to these three degrees of dependence: rules with $Pr(A \cup C, \mathcal{D}) > Pr(A, \mathcal{D}) \times Pr(C, \mathcal{D})$ are called the *positive association rules*, those with $Pr(A \cup C, \mathcal{D}) < Pr(A, \mathcal{D}) \times Pr(C, \mathcal{D})$ are the *negative association rules* and finally, $Pr(A \cup C, \mathcal{D}) = Pr(A, \mathcal{D}) \times Pr(C, \mathcal{D})$ are *null association rules*.

[1] This work is supported in part by MCYT TIC 2002-04019-C03-01
[2] Universitat Politècnica de Catalunya, Barcelona, email: gcasas@lsi.upc.edu

A well-known measure that evaluates the degree of dependency between antecedent and consequent of a rule is the Pearson coefficient, noted by ϕ. Rules with $\phi = 0$ are independent, rules with $\phi > 0$ are the positive rules and the rest with $\phi < 0$ are the negative rules. So, to check independence between two variables (in our case antecedent and consequent of a rule) we could perform the common statistical correlation testing by rejecting or accepting the hypothesis **H0)** $\phi = 0$, versus **H1)** $\phi \neq 0$ (the convenient transformation of ϕ gets a statistic that follows normality). Unfortunately, Pearson coefficient fails to fulfill property P2.

The second predicate P2 illustrates the need to distinguish the best association rules from all the antecedent-consequent permutation asymmetries. All the rules r whose value $IM(r) < IM(\hat{r})$ are said to be a *weak reverse of another rule*.

We can finally define our working hypothesis for which an interestingness measure IM is accurate if it can prune misleading rules, i.e, **weak rules** (null association rules and negative association rules) and **weak reversed rules**. Null association rules are useless since we are looking for association patterns and not independent ones; and we consider that negative association rules should be better discovered with different specific algorithmic strategies having into account the negation of attributes, such as in [8], where the necessary monotonicity properties are preserved. This total set of rules that IM has to prune will be called the **uninteresting rules**.

2.1 Useful Tests on Rules

The prune phase becomes a rule classification problem currently performed through the ranking set up by IM. We formalize it as a test $\mathbf{T}$ on an association rule r: from the input database $\mathcal{D}$, given an interestingness measure IM and a certain threshold θ a test is,

$$\mathbf{T}(r, IM, \theta, \mathcal{D}) = \begin{cases} 1 & \text{if } (IM(r,\mathcal{D}) > \theta) \text{ and} \\ & \quad (IM(r,\mathcal{D}) > IM(\hat{r},\mathcal{D})), \\ 0 & \text{otherwise.} \end{cases}$$

When this test returns 1 means that the association rule r is considered interesting in the concrete database $\mathcal{D}$, otherwise it should be pruned away. In a certain way, the first part of the condition, $IM(r,\mathcal{D}) > \theta$, controls the satisfactibility of property P1; and the second part, $IM(r,\mathcal{D}) > IM(\hat{r},\mathcal{D})$, controls the satisfactibility of property P2. Of course, the utility of the test and the ability to capture interestingness depends basically on IM and the value of θ chosen.

A test will be considered **harmless** if all the interesting rules pass the test, although it could still hold many uninteresting rules at the same time. We say it is harmless because at least interesting rules are *never* removed.

A test will be considered **completely useful** if it perfectly separates uninteresting rules from the rest, so, it always performs a perfect classification and never fails to distinguish the notion of interestingness. Any completely useful test is included in the set of harmless tests, but the reverse implication does not always hold. For our goals, we want to consider only all the completely useful tests.

2.2 Partial Orders on Rules

We propose to study the following three partial orders on rules.

Definition 2.1 *Given rules $r = A \Rightarrow C$, and $r' = A' \Rightarrow C'$, we say $r <_1 r'$ in a certain database $\mathcal{D}$ if and only if: $Pr(A,\mathcal{D}) = Pr(A',\mathcal{D})$,* **and** *$Pr(C,\mathcal{D}) = Pr(C',\mathcal{D})$,* **and** *$Pr(A \cup C,\mathcal{D}) < Pr(A' \cup C',\mathcal{D})$.*

Definition 2.2 *Given rules $r = A \Rightarrow C$, and $r' = A' \Rightarrow C'$, we say $r <_2 r'$ in a certain database $\mathcal{D}$ if and only if: $Pr(A \cup C,\mathcal{D}) = Pr(A' \cup C',\mathcal{D})$,* **and** *$Pr(C,\mathcal{D}) = Pr(C',\mathcal{D})$,* **and** *$Pr(A,\mathcal{D}) < Pr(A',\mathcal{D})$.*

These two partial orders on rules derive from the well-known properties proposed by Piatetsky-Shapiro [11] over the measures of interestingness.

Definition 2.3 *Given rules $r = A \Rightarrow C$, and $r' = A' \Rightarrow C'$, we say $r <_3 r'$ in a certain database $\mathcal{D}$ if and only if: $Pr(A \cup C,\mathcal{D}) = Pr(\overline{A'} \cup \overline{C'},\mathcal{D})$,* **and** *$Pr(\overline{A} \cup \overline{C},\mathcal{D}) = Pr(A' \cup C',\mathcal{D})$,* **and** *$Pr(\overline{A} \cup C,\mathcal{D}) = Pr(A' \cup \overline{C'},\mathcal{D})$,* **and** *$Pr(A \cup \overline{C},\mathcal{D}) = Pr(\overline{A'} \cup C',\mathcal{D})$,* **and** *$Pr(A \cup C,\mathcal{D}) < Pr(A' \cup C',\mathcal{D})$ (where $\overline{X}$ means the absence of itemset X in the database $\mathcal{D}$).*

This third partial order on rules expresses the relationship that should exist between two complementary rules: that is, rules that would have the same support in case all the 1's (presence of item in a transaction) would be flipped into 0's (absence of item) simultaneously in all transactions of $\mathcal{D}$. So, the order of $<_3$ reflects that the co-presence of antecedent and consequent in each transaction is more meaningful that their co-absence.

From these three partial orders, we define a total proper order that measures IM should keep to rank the rules. Later, we will show that some total orders induced by specific measures, we have that they are proper orders.

Definition 2.4 *A measure IM induces a proper order if preserves the partial orders $<_1$, $<_2$ and $<_3$ given in $\mathcal{D}$. That is, $r <_1 r'$ or $r <_2 r'$ or $r <_3 r' \longrightarrow IM(r) \leq IM(r')$*

3 PROPERTIES OF AN OPTIMAL PRUNE

According to our framework, the main goal of an optimal prune is to find a completely useful test with the ability to keep a proper order on those interesting surviving rules. To find a completely useful test we are going to consider symmetry breaking measures IM (this excludes ϕ, and also other current mesures like Lift, PS [11] or IS[13]), and analyze how the chosen threshold θ affects the properties of the measure IM.

We start by observing that given any symmetry breaking IM, it is always possible to find a threshold θ that makes the test $\mathbf{T}(r, IM, \theta, \mathcal{D})$ harmless. This can be done by setting the threshold θ with the smallest value of the image IM, that is, if $IM(r,\mathcal{D}) \in [v_s, v_e]$, then we can chose $\theta = v_s$, so that the test *always* returns 1 (all the rules pass the test). The problem of using this minimum harmless threshold is that the test is not useful at all because all the uninteresting rules still hold. The point is how well we can do with θ, i.e. how much we can increment the value of θ keeping the test $\mathbf{T}(r, IM, \theta, \mathcal{D})$ still harmless and, at the same time, with the ability to remove uniteresting rules.

Definition 3.1 *The* maximum harmless threshold*, noted by θ^*, for some symmetry breaking measure IM is that value for θ such that if we incremented this value θ^* with any $\delta > 0$, then the test $\mathbf{T}(r, IM, \theta^* + \delta, \mathcal{D})$ would start being harmful.*

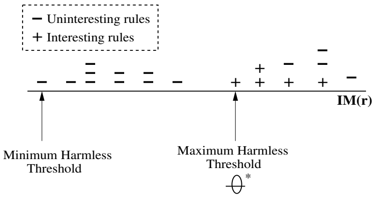


Figure 1. Dotplot of a certain $IM(r)$ where θ^* leads to a harmless test

The definition of θ^* removes as many uninteresting rules as possible but it always keeps the harmless condition of the test. A graphical example of the threshold θ^* for a measure IM is found in figure 1. This graph shows a dotplot of $IM(r)$: location of the points ($+$ and $-$) along the line $IM(r)$ represents the different values that each rule gets with IM. As we see, interesting and uninteresting rules could be mixed along the line, but at least, the threshold θ^* guarantees a set of *only* uninteresting rules at its left side, and it cannot be incremented to hold this invariant.

Proposition 3.1 *The value of the maximum harmless threshold for any IM is $\theta^* = \min_{r_i} \{IM(r_i, \mathcal{D})\}$ where r_i are the interesting rules found in the data $\mathcal{D}$.*

With $\theta^* = \min_{r_i} \{IM(r_i, \mathcal{D})\}$ we are in the limit of harmlessness in a test, holding as few uninteresting rules as possible in the right side of θ^*. These uninteresting rules, noted by r_u, are weak rules or weak reversed rules, but they still could pass the test if and only if IM is not reliable, that is, $IM(r_u, \mathcal{D}) \geq IM(r_i, \mathcal{D})$ for some interesting rule r_i. Since the test is harmless, it cannot remove r_i, and the following situation is forced: $IM(r_u, \mathcal{D}) \geq IM(r_i, \mathcal{D}) \geq \theta^*$. The number of r_u kept with the maximum harmless threshold is minimum by definition; but, when can this maximum harmless threshold perform a perfect classification of rules?

Proposition 3.2 *The maximum harmless threshold performing a perfect split of interestingness, exists for any symmetry breaking IM if we have that $\max_{r_u} \{IM(r_u, \mathcal{D})\} < \min_{r_i} \{IM(r_i, \mathcal{D})\}$, where r_u are the uninteresting rules and r_i are the interesting ones in the data $\mathcal{D}$.*

This threshold θ^* will convert the test in completely useful when all the rules r_u are removed by the test. This situation only happens when we have that $\max_{r_u} \{IM(r_u, \mathcal{D})\} < \min_{r_i} \{IM(r_i, \mathcal{D})\}$, and so we can choose θ^* such that $\forall r_u \; IM(r_u, \mathcal{D}) < \theta^*$, but at the same time, $\forall r_i IM(r_i, \mathcal{D}) \geq \theta^*$. In other words, the function IM assigns values to rules in such a way that iteresting rules r_i are separated from the rest of uninteresting rules r_u (and the corresponding split between these two type of rules is pointed out by θ^*). However, the existence of a θ^* for IM giving rise to a completely useful test, depends on the specific data examined $\mathcal{D}$ and, specially, on the ability of the measure to clearly separate the two type of rules at this point θ^*. In particular, we can state the followig.

Lemma 3.1 *If a certain symmetry breaking IM is linearly correlated with ϕ, then $\exists\theta^*$ creating a completely useful test $\mathbf{T}(r, IM, \theta^*, \mathcal{D})$.*

Proof. Given the input set of all-kind rules R to be classified, we can construct a new set R' consisting of only the strong reversed rules, i.e, $R' = \{r \in R | IM(r) > IM(\hat{r})\}$ (this can be done because our IM is symmetry breaking). Besides, if IM is linearly correlated with ϕ, it implies that the relationship can be graphically induced by a line $IM = m \cdot \phi + b$ (where m is the slope and b is the y-intersection point of the line). This linear dependence allows to set a *single* cut-off point from where to start distinguishing strong positive rules from the rest of weak rules (from the cut-off value of ϕ). This point gives a single threshold θ as a classification point for IM in the test. That is, we can create from set R' a partition such that $\max_{r_w} \{IM(r_w, \mathcal{D})\} < \min_{r_s} \{IM(r_s, \mathcal{D})\}$, where r_w are the weak rules in R' and r_s are the strong positive association rules in R'. But since R' just contained strong reversed rules, we have that rules r_s are also the interesting ones (strongly correlated and the strong reversed ones). □

Moreover, we want the measure IM to induce the proper order on the remaining interesting rules. Lemma 3.2 states the three necessary conditions for a measure to establish a proper order. Finally, table 1 gathers a comparison of the some current measures (see [13] for formulas) according to these three conditions.

Lemma 3.2 *For all rules $r = A \Rightarrow B$, the following conditions, taken jointly, are sufficient for establishing that a total order induced by IM is a proper order:*

1. *$IM(r, \mathcal{D})$ is monotone in $Pr(A \cup C, \mathcal{D})$ over rules with the same $Pr(A, \mathcal{D})$ and same $Pr(C, \mathcal{D})$.*
2. *$IM(r, \mathcal{D})$ is monotone in $Pr(A, \mathcal{D})$ over rules with the same $Pr(A \cup C, \mathcal{D})$ and same $Pr(C, \mathcal{D})$.*
3. *$IM(r, \mathcal{D})$ is monotone in $Pr(A \cup C, \mathcal{D})$ over complementary rules.*

IM	(1)	(2)	(3)	IM	(1)	(2)	(3)
ϕ	Yes	Yes	Yes	$\tilde{\phi}$	Yes	Yes	Yes
Confidence [1]	Yes	Yes	Yes	Lift	Yes	Yes	Yes
Conviction [6]	Yes	Yes	Yes	PS [11]	No	Yes	Yes
Gini Index	No	No	No	IS [13]	Yes	Yes	Yes
Inf. Gain	No	No	Yes	J-Measure	Yes	No	No

No, unless only positive association rules are considered.

Table 1. Conditions of lemma 3.2 satisfied by main IM

Note that in this table we have already introduced a new meassure, $\tilde{\phi}$; this will be defined later as a part of our proposal.

4 MULTI-TEST APPROACH

To find the completely useful test, we study current symmetry breaking measures able to induce a final proper order (Confidence, Conviction and J-Measure). This study will show the necessity to later introduce our new measure, $\tilde{\phi}$. For comparison purposes, we generate artificial datasets such as in [13] containing $10,000$ random samples. Each sample is a 2×2 contingency table representing an association rule $X \Rightarrow Y$. Each generated contingency table is subject to the same restrictions as in [13] and a given minimum support σ will represent the support-based prune performed by the frequent sets algorithms on the first phase. In the following synthetic experiments we assume $\sigma = 0$ (all the possible rules are generated).

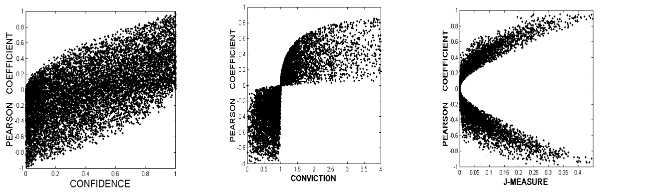


Figure 2. Correlation of Confidence, Conviction and J-Measure against ϕ

At a glance, comparisons of these main symmetry breaking measures to ϕ can be grasped from figure 2 (we compare it to ϕ due to lemma 3.1). Note that the interesting rules are exactly located in the high top half of each square, that is, those with $\phi >> 0$ and with no other stronger reverse. We realize that none of these measures can perform a perfect prune of all uninteresting rules at a time. A test can be regarded as a split along the vertical line $y = \theta$, and whatever the threshold θ chosen, the test $\mathbf{T}(r, IM, \theta, \mathcal{D})$ will always keep null or negative rules; thus, proposition 3.2 never holds.

Of course, another possibility before proposing our new measure that performs a single prune, is to try to combine different measures to create a multi-test proposal achieving the three goals of a completely useful test: 1) pruning null ass. rules, 2) pruning negative ass. rules, 3) pruning weak reversed rules. For example, $\mathbf{T}(r, \phi, \theta_1, \mathcal{D})$ *and* $\mathbf{T}(r, Conviction, 1, \mathcal{D})$ is a completely useful multi-test: ϕ with a convenient threshold θ_1, keeps only the strongest rules; and then, those rules go to the second test where Conviction with a harmless threshold will prune the worst reversed rules and keep the proper order on the rest. Note that the threshold θ_1 for the measure ϕ could be determined statistically by studying the distribution function of ϕ. More complex combinations can be done: for instance, $\mathbf{T}(r, Conviction, 1, \mathcal{D})$ *and* $\mathbf{T}(r, J - Measure, \theta_2, \mathcal{D})$.

5 A NEW MEASURE FOR AN OPTIMAL PRUNE

We study here a perfect IM: it should be symmetry breaking (P2), it should be able to remove null and negative rules (P1), and keep a final proper order. For that, we observe that the Pearson coefficient ϕ just fails to fullfil predicate P2; so, the most natural approach seems to modify ϕ and transform it into a symmetry breaking measure.

In general, when examining association rules, we should take into account that the *best* rule in terms of *implication*, $A \Rightarrow C$, comes when the transactions where antecedent A occurs are a subset of the transactions where consequent C occurs (i.e, $trans(A, \mathcal{D}) \subseteq trans(C, \mathcal{D})$). In other words, the occurrence of A in the database fully implies the occurrence of C. Besides, transactions where A occurs can be divided into the following: $trans(A, \mathcal{D}) = trans(A \cup C, \mathcal{D}) + trans(A \cup \neg C, \mathcal{D})$. So, the fewer transactions in which $A \cup \neg C$ occurs, the better for the rule $A \Rightarrow C$ (this implies that the support of A is mainly due to $A \cup C$ where both itemsets occur together, and we get closer to the inclusion $trans(A, \mathcal{D}) \subseteq trans(C, \mathcal{D})$).

To incorporate this reasoning in the Pearson coefficient ϕ, we examine the contingency table from where its value is calculated. Given two itemsets X and Y, we study the counting supports for the occurence of one variable without the other, and viceversa, and we can conclude that:

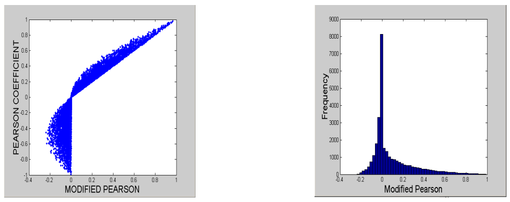
- If $trans(X \cup \neg Y, \mathcal{D}) > trans(\neg X \cup Y, \mathcal{D})$, we would choose the implication $Y \Rightarrow X$.
- If $trans(X \cup \neg Y, \mathcal{D}) < trans(\neg X \cup Y, \mathcal{D})$, we would choose the implication $X \Rightarrow Y$.

For a general rule $A \Rightarrow C$, these two observations can be expressed by the ratio $\frac{Pr(A \cup C, \mathcal{D})}{Pr(A, \mathcal{D})}$, that is, the bigger proportion of the antecedent that is shared with the consequent, the better. The easiest way to modify the Pearson coefficient ϕ to incorporate this knowledge without losing the ability to prune weak rules, is then the following:

$$\tilde{\phi}(A \Rightarrow C) = \phi(A, C) \times \frac{Pr(A \cup C, \mathcal{D})}{Pr(A, \mathcal{D})}$$

i.e, the product of confidence of the rule times its Pearson coefficient. We note that confidence forms part of the well-known framework that states that strong rules have support and confidence over a user-specified threshold ([3]); this makes our measure also suitable for that framework, but even solving some of the common inconvenients. In particular, the inconvenient of confidence (see [6] or [4]) is that independent rules $r = A \Rightarrow C$ have a confidence equal to $Pr(C, \mathcal{D})$, which could be still high enough to make the rule hold, and only positive association rules have confidence over $Pr(C, \mathcal{D})$. However, this lack of variability in the presence of the consequent in the data does not let us to be sure about the rule. With our measure, this problem is solved: we know by construction that if a rule r is independent then $\tilde{\phi}(r) = 0$, regardless of the value for confidence; and if the rule is positive dependent then $\tilde{\phi}(r) > 0$ since confidence can never have a negative value. On the other hand, the new measure $\tilde{\phi}$ can be also regarded a transformation of ϕ that gets to be symmetry breaking: $\tilde{\phi}(r) > \tilde{\phi}(\hat{r})$ if $confidence(r) > confidence(\hat{r})$.

In figure 3 we see the behaviour of this measure for synthetic data: $\tilde{\phi}$ is highly correlated with ϕ for positive rules (see figure 3), even keeping almost the same scale (this is good to distinguish strong positive rules from the weak rules as pointed out by lemma 3.1); so, $\tilde{\phi}$ distinguishes positive association rules from the rest and also it is symmetry breaking. Values of $\tilde{\phi}$ for negative dependent rules have more variability; however, this value of $\tilde{\phi}$ for negative rules will never be over zero, which eases the optimal prune. In conclusion, $\tilde{\phi}$ will be correlated with ϕ for positive dependent rules, and the value of zero gives us a point from where to start pruning in a harmless way. Moreover, the new measure induces still a proper order.



$\tilde{\phi}$ against ϕ Histogram $\tilde{\phi}$

Figure 3. Behaviour of our proposal $\tilde{\phi}$ with synthetic uncorrelated data

5.1 Maximum Harmless Threshold for $\tilde{\phi}$

We know that a symmetry breaking IM with the ability to prune weak and null rules, can potentially construct a completely useful test. However, this will depend on the value for the threshold θ^*, that should represent a perfect split between interesting rules and uninteresting rules (see proposition 3.1 and 3.2).

Threshold θ^* only plays a role on the first part of the condition of the test (to decide if the antecedent and consequent are correlated according to ϕ). Hence, to approach the study of this harmless value of θ^* that creates a completely useful test, we study the acceptance or rejection of the following hypothesis: **H0)** r **is an uncorrelated association rule** ($\tilde{\phi} \leq 0$) versus **H1)** r **is strong positive association rule** ($\tilde{\phi} > 0$). The cut-off point that distiguishes these two hypothesis at a certain user-specified significance level will give the value we want for θ^* (hypothesis testing through a cut-off point and a p-value, is a common technique used in statistics).

So, we now study the distribution function of θ^* for uncorrelated data (i.e, under the hypothesis H0). In figure 3 we see that the histogram of $\tilde{\phi}$ for this kind of data does not follow normality; so, the probability density function of the new measure, and so, its distribution function, can be difficult to approximate theoretically. In this paper we will use as an approximation the empirical distribution function of a sequence of realizations of $\tilde{\phi}$ for randomly-generated rules. That is, if $\phi^* \sim f$, and $x_1, \ldots, x_n$ is a sample for values of ϕ^*, then we approximate $\hat{f}_n$ with this sample (the well-known theorem by Glivenko-Cantelli ensures this is a good way to aproximate the real distribution function as the sample size becomes bigger).

Sample Size	Cut-off at 99%	Cut-off at 95%
130,000	0.7700	0.5302
140,000	0.7696	0.5302
150,000	0.7694	0.5309
160,000	0.7682	0.5308
170,000	0.7682	0.5308
180,000	0.7682	0.5308

Table 2. Simulation of empirical distribution of $\tilde{\phi}$

Simulation of different samples will lead to a good approximation of the real distribution function, and we will be able to infer from there the cut-off point at the significance levels of 99% and 95%. Table 2 shows the different simulations and results for growing samples. As the sample becomes bigger, the cut-off points become more stable. Finally, we decide to take as inferred value $\theta^* = 0.7682$ to determine the statistically significant interestingness of rules at a level of 99%, and $\theta^* = 0.5308$ at a level of 95%. Other methods to infer the density function, and from there the distribution function, could have been applied: for instance kernel methods of non-parametrics statistics, or fitting a Johnson curve to find the exact formula.

6 EXPERIMENTS AND CONCLUSIONS

We follow the following one-step strategy: Order by $\tilde{\phi}$ those rules r such that $\tilde{\phi}(r) > 0.7682$.

Since $\tilde{\phi}$ induces a proper order, no more than one single step is needed to prune all the uninteresting rules. For synthetic data we generated synthetic 10,000 initial association rules such as in [13], considering that the minimum support threshold is $\sigma = 0$, so all the possible rules are generated. With just one step the strategy removes all the uninteresting rules keeping just 113 final rules, that have a confidence over 99%. So, these are the stronger ones.

The next goal is to perform tests using real databases. We used a sample of the USA census from PUMS[3] consisting of 3000-transaction database of 80 possible items. In contrast with synthetic experiments, we used now a $\sigma = 0.15$ and we got a total of 26,164 initial association rules. These total rules are ploted in figure 4. After the prune, there is only 142 surviving rules; all of them turn to have a confidence over 99%. It is worth noting that our proposed measure is objective and it does not take into account any subjective considerations. Thus, once the strongest patterns are separated from the rest, the user can use other subjective measures of interestingness over the remaining rules (see [12]).

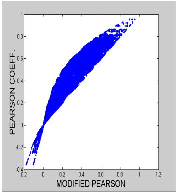


Figure 4. Behaviour of the new measure with real data

In this paper we formalize the optimal prune of association rules with a completely useful test created by a maximal harmless threshold, that is, a test with a measure IM with properties P1 and P2 and keeping a proper order on rules. This formalization allows the evaluation of current different measures. We also present a new measure, $\tilde{\phi}$, that meets all the necessary requirements for the optimal prune. More experiments should be performed.

[3] www.ipums.umn.edu/usa/intro.html

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. *Proc. Int'l Conf. on Management of Data*, 207-216. 1993.

[2] Y.Bastide, N.Pasquier, R.Taouil, G.Stumme and L.Lakhal. Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets. *First Int'l Conf. on Computational Logic*, 972-986. 2000.

[3] R.Bayardo Jr, and R.Agrawal. Mining the Most Interesting Rules. *Proc. Int'l Conf. on Knowledge Discovery and Data Mining*, 145-154. 1999.

[4] F.Berzal, I.Blanco, D.Sánchez and María-Amparo Vila Measuring the Accuracy and Interest of Association Rules: A new Framework. *Journal Intelligent Data Analysis*, Vol 6, pages 221-235. 2002.

[5] S.Brin, R.Motwani, and C.Silverstein. Beyond market baskets: Generalizing association rules to correlations. *Proc. Int'l Conf. on the Management of Data*, 265-276. 1997.

[6] S.Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proc. Int'l Conf. on Management of Data*, Volume 6:2, 255 - 264.1997.

[7] L.Cristofor, and D.Simovici. Genearating an Informative Cover for Association Rules. *Int'l Conf. on Data Mining*, p.597. 2002.

[8] I.Fortes, J.L.Balcázar, and R.Morales. Bounding Negative Information in Frequent Sets Algorithms. *Int'l Conference on Discovery Science*, 50-58, 2001.

[9] R.Hilderman, and H.Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 247 - 259. 2001.

[10] B.Liu, W.Hsu, and Y.Ma. Pruning and Summarizing the Discovered Associations. *Proc. Int'l Conf. on Knowledge Discovery and Data Mining*, 125-134. 1999.

[11] G.Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. *Int'l Conf. on Knowledge Discovery in Databases*, 229-248. 1991.

[12] A.Silberschatz, and A.Tuzhilin. On Subjective Measures of Interestingness in Knowledge Discovery. *Proc. Int'l Conf. Knowledge Discovery in Databases*, 275-281. 1995.

[13] P.Tan, and V.Kumar. Interestingness Measures for Association Patterns: A Perspective. *KDD Workshop on Postprocessing in Machine Learning and Data Mining*. 2000.