Statistical Strategies for Pruning All the Uninteresting Association Rules *

G. Casas-Garriga
Departament de LSI
Universitat Politècnica de Catalunya
gcasas@lsi.upc.es

April 11, 2003

Abstract

We propose a general framework to describe formally 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. Final experiments serve to show the ability of our proposal.

1. Problem Formulation and Basic Definitions

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 $\mathcal{I} = \{i_1, i_2, \dots, i_N\}$. 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 the previous 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 are 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 (such as in [4], [6], [5], [11], [14], [16] ...), as opposed to other deterministic techniques such as grouping together related rules ([8]), or using closed itemsets to generate a final non-redundant set of rules ([3] or [18]).

We introduce now some definitions and consideration in our problem.

*This work is supported in part by EU ESPRIT IST-1999-14186 (ALCOM-FT), and MCYT TIC 2002-04019-C03-01 (MOISES)

Definition 1.1 *An* interestingness measure *IM is a function on association rules that returns a real value, that is, $IM : \{Association\ Rule, \mathcal{D}\} \longrightarrow \Re$.*

So, interestingness measures aim at sorting association rules according to this output real value. 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 function IM 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 need to consider an asymmetric framework where one variable causes another. So, there is a need to distinguish the strenght 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*.

Definition 1.2 *We say that the association rule $r = X \Rightarrow Y$ is a* better implication *in a database $\mathcal{D}$ than its reversed $\hat{r} = Y \Rightarrow X$, according to a measure IM, if $IM(r, \mathcal{D}) > IM(\hat{r}, \mathcal{D})$.*

2. General Framework for Pruning Association Rules

This following proposed framework tries to be a generalization of all the different properties and considerations stated in the broad current literature ([10], [16],[17], [14],[6],[11],[13], among others).

2.1. Necessary properties for Interestingness Measures

The proposed properties stem from intuitive notions of interestingness considered from an objective point of view in the context of the association rule mining. Given any interestingness measure IM, we consider two properties making IM an accurate metric in the assessment of association rules:

P1: IM must test independence of a rule r

P2: IM must test the strenght 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 [5] argue, this fact is true to some extent because itemsets with high support are a source of misleading rules: they appaear 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. For example, adding a new item i' to $\mathcal{I}$ and including it in the transactions of the database, so that i' appears in all the transactions, gives rise to frequent itemsets where i' is always present. However, when generating the subsequent rules, most of them 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 says that any accuracy measure must test independence between antecedent and consequent of a rule. Stated 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 [14]. 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*.

A well-known measure that evaluates the degree of dependency between antecedent and consequent of a rule is the Pearson coefficient, ϕ (see appendix A for more details). 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 an statistic that follows normality). Unfortunately, Pearson coefficient fails to fulfill property P2; so, it is not a good measure to be used in the association rule mining framework and other measures should be considered.

The second predicate illustrates the need to distinguish the best association rules from all the antecedent-consequent permutation asymmetries. In other words, given that we are in the asymmetric framework of association rules, we just want to keep one single representative from any pair of rules r and $\hat{r}$. 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 [9], where the necessary monotonicity properties are preserved, which is not necessarily the case for statistical metrics [13]. This total set of rules that IM has to prune will be called the **uninteresting rules**.

2.2. Useful Tests on Rules

The last prune phase becomes a rule classification problem that is currently performed through the ranking stablished by IM. It can be formalized through the following test.

Definition 2.1 *A test* **T** *on an association rule* r *from the input database* $\mathcal{D}$*, given an interestingness measure* IM *and a certain threshold* θ *is:*

```
$\mathbf{T}(r, IM, \theta, \mathcal{D}) =$
    if  ( $IM(r,\mathcal{D}) > \theta$  and  $IM(r,\mathcal{D}) > IM(\hat{r},\mathcal{D})$ ),   then return 1
    otherwise return 0
```

When this test returns 1 means that the association rule r is considered interesting in the concrete database $\mathcal{D}$, otherwise, returning a 0, it means that r is not considered interesting and it should be pruned away. In a certain way, if we examine closely the main condition of the test, we note that the first part, $IM(r,\mathcal{D}) > \theta$, controls the satisfactibility of property P1; and the second part, $IM(r,\mathcal{D}) > IM(\hat{r},\mathcal{D})$, controls the satisfactivility of property P2. Of course, the utility of the test depends basically on IM and the value of θ chosen (that will determine the ability of the test to capture interestingness). We want to distinguish here two degrees of ability in a test.

A test will be considered **harmless** if all the real 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 real 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 of rules 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 (i.e, there are harmless tests which are not completely useful). For our goals, we want to consider only all the completely useful tests, altough this will depend on IM and the threshold θ used as a cut-off.

2.3. Partial Orders on Rules

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

Definition 2.2 *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 2.3 *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 [14] over the measures of interestingness.

Definition 2.4 *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. In other words, in the market basket framework, the antecedent and the consequent should be strongly associated if they are bought together by many costumers, rather than because they are not bought together frequently.

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 2.5 *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 $r <_3 r' \longrightarrow IM(r) \leq IM(r')$*

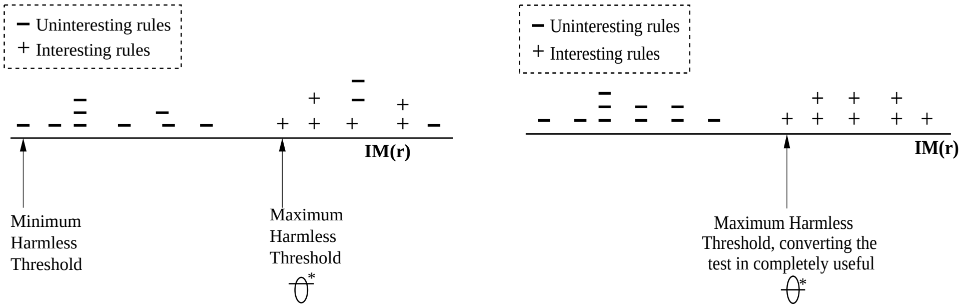
3. Determining the Properties of an Optimal Prune

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. For that, we focus our study on how the chosen threshold θ affects the properties of the measure IM.

3.1. Finding a Completely Useful Test

We are going to consider here symmetry breaking measures IM (this excludes ϕ, that can never lead to a optimal prune due to P2), and analyze which characteristics the value of θ must fulfill to create a completely useful, whenever this is possible.

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 $\theta = v_s$. This will always make the test $\mathbf{T}(r, IM, v_s, \mathcal{D})$ harmless since *always* returns 1, and so, all the rules pass the test. This naive value of θ will be called the *minimum harmless threshold* of IM.


− Uninteresting rules
+ Interesting rules
IM(r)
Minimum Harmless Threshold
Maximum Harmless Threshold
θ*
− Uninteresting rules
+ Interesting rules
IM(r)
Maximum Harmless Threshold, converting the test in completely useful
θ*

(a) θ^* leads to a harmless test (b) θ^* leads to a completely useful test

Figure 1: Dotplots of values $IM(r)$ to illustrate different types of test

The problem with using the minimum harmless threshold is that the test is not useful at all because all the uninteresting association rules are kept. So, the point is how well we can do with θ, i.e, how much we can increment the value of θ keeping the test $\mathbf{T}(r, IM, \theta, \mathcal{D})$ being 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 a certain δ, then the test $\mathbf{T}(r, IM, \theta^* + \delta, \mathcal{D})$ would start being harmful.*

So, θ^* removes as many uninteresting rules as possible, but always keeps the harmless condition of the test. A graphical example of the threshold θ^* for a measure IM, is found in graph (a) of figure 1. This graph shows a dotplot of $IM(r)$: location of the points (+ and −) along the line $IM(r)$ shows 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^* always 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 $IM(r_u, \mathcal{D}) \geq IM(r_i, \mathcal{D})$, for some interesting rule r_i. Since the test is harmless, it cannot remove r_i, and the following situation is forced: $IM(r_u, \mathcal{D}) \geq IM(r_i, \mathcal{D}) \geq \theta^*$. So, r_u is an uninteresting rule that the test is forced to keep just to continue being harmless. However, we can state that 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 interesting 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^*). Graph (b) of figure 1 shows the situation of proposition 3.2. However, the existence of a θ^* for IM giving rise to a completely useful test, depends on the especific data examined $\mathcal{D}$ and especially, 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 the test $\mathbf{T}(r, IM, \theta^*, \mathcal{D})$ in completely useful.*

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 IM can distinguish strong positive rules from the rest of weak rules. 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 association rules in R' and r_s are the strong 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). So, this IM can separate rules acccording to proposition 3.2, which implies that the maximum harmless threshold converting the test in completely useful exists for IM. □

3.2. Keeping a Proper Order on Rules

Besides, this measure IM used in the test should induce a proper order on the remaining interesting rules. The table 1 gathers the conditions satisfied by the different measures (see appendix B for definitions). Note that measures like Lift, PS or IS will never create the completely useful test since they are not symmetry breaking. Measures not inducing a proper should be also discarded.

Lemma 3.2 *For all rules $r = A \Rightarrow B$, the following conditions, taken joinly, 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.

Proof sketch: As appendix C.

Table 1: **Conditions satisfied by main** IM

IM	**(1)**	**(2)**	**(3)**	**IM**	**(1)**	**(2)**	**(3)**
ϕ	Yes	Yes	Yes	$\tilde{\phi}$	Yes	Yes	Yes
Confidence	Yes	Yes	Yes	Lift	Yes	Yes	Yes
Conviction	Yes	Yes	Yes	PS [14]	No	Yes	Yes
Gini Index	No	No[1]	No[1]	IS [16]	Yes	Yes	Yes
Inf. Gain	No	No[1]	Yes	J-Measure	Yes	No[1]	No[1]

[1] No, unless only positive association rules are considered.

4. Multi-test Approach

To find the completely useful test, the current symmetry breaking measures also able to induce a final proper order should be studied (Confidence, Conviction [7] and J-Measure). For comparison purposes, we generate artificial datasets such as in [16] containing 10,000 random samples. Each

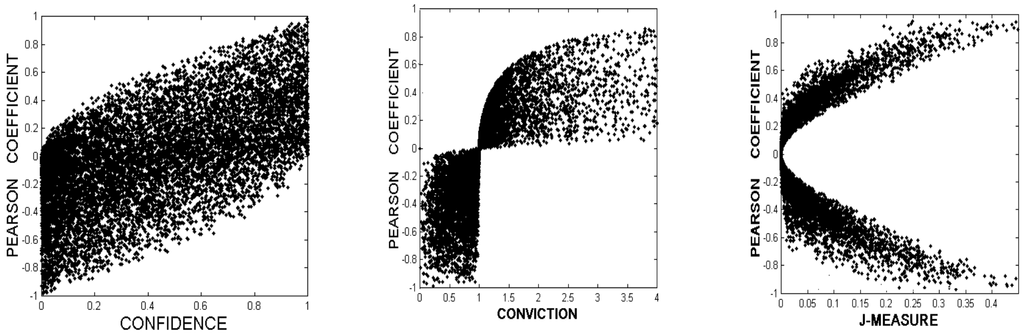


Figure 2: Correlation of Confidence, Conviction and J-Measure against ϕ

sample is a 2×2 contingency table representing an association rule $X \Rightarrow Y$ where $X, Y \subset \mathcal{I}$. Each generated contingency table is subject to the same restrictions as in [16]. Apart from these restrictions, a 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 $\sigma = 0$, i.e, we are dealing with the worst case where all the possible rules are generated.

At a glance, comparisons of this main symmetry breaking measures to ϕ can be grasped from figure 2. Note that the interesting rules we want to keep are exactly located in the high top half of each square, that is, those with $\phi >> 0$ and with no other stronger reverse (which are not plotted in these figures). From the graphics we can see 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 for these measures, the test $\mathbf{T}(r, IM, \theta, \mathcal{D})$ will always maintain null association rules or negative association rules; thus, the proposition 3.2 never holds.

However, although a completely useful test is not possible with one single measure, we can try to combine them to create a multi-test proposal achieving the three goals of a completely useful test: 1) pruning null association rules, 2) pruning negative association 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 the 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 threhold θ_1 for the measure ϕ could be determined statistically by studying the distribution function of ϕ (in the same way that one can perform a correlation test to decide the significance of ϕ between two variables).

More complex combinations can be done: $\mathbf{T}(r, Conviction, 1, \mathcal{D})$ *and* $\mathbf{T}(r, J-Measure, \theta_2, \mathcal{D})$. Here, Conviction with this harmless threshold, $\theta = 1$, prunes all the negative association rules (see figure 2) and all the weak reversed rules. Finally, J-Measure in the second test would prune all the null association rules. The harmless value of θ_2 for J-Measure is here more difficult to determine theorethically (from figure 2 we see that θ_2 is somewhere around 0.25).

5. A New Measure to Have an Optimal Prune

We want to study now the existance of a perfect IM: it should be symmetry breaking (P2), it should be able remove null and negative rules (P1), and keep a final proper order. This single measure could certainly make the post-prunning phase faster and simpler, since just one single statement should be checked for each association rule. For that, we observe that the Pearson coefficient ϕ just fails to fullfil predicate P2; so, the most natural approach to this problem 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, $trans(A) \subseteq trans(C)$). 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: $trans(A) = trans(A \cup C) + trans(A \cup \neg C)$. 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 $trans(A) \subseteq trans(C)$).

To incorporate this reasoning in the Pearson coefficient ϕ, we examine the contingency table from where its value is calculated (see table 3 in appendix A). Given two itemsets X and Y, we study the values f_{01} and f_{10} (i.e, counting supports for the occurence of one varible without the other, and viceversa), and we can conclude that:

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

For a general rule $A \Rightarrow C$, these two observations can expressed by the ratio $\frac{Pr(A \cup C)}{Pr(A)}$, that is, the bigger proportion of the antecedent that is shared with the consequent, the better. Or in other words, the ratio gives the strenght of the implication in case we chose A as antecedent. 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:

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

i.e, the product of confidence of the rule times its Pearson coefficient (definition of the Pearson coefficient in appendix A). We note that confidence forms part of the well-known framework that states that strong rules have support and confidence over the user-specified threshold ([4]); this makes our mesure also suitable for that framework, but even solving some of the inconvenients that have been stated in the current literature.

In particular, the inconvenient of confidence (see [7] or [5]) 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 allow 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.

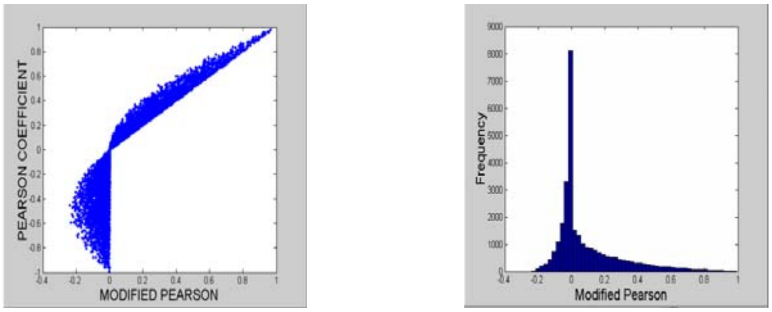
Values of $\tilde{\phi}$ for negative dependent rules have more variability. However, this value of $\tilde{\phi}$ for negative association rules will never be over zero, which eases the optimal prune. In other words, $\tilde{\phi}$ will be correlated with ϕ for positive dependent rules, and the value of zero give us a point from where to start pruning in a harmless way.

Apart from that, te new measure $\tilde{\phi}$ can be regarded a transformation of ϕ that gets to be symmetry breaking; so, it can distinguish the strenght of both implications. We know that $\tilde{\phi}(r) > \tilde{\phi}(\hat{r})$ if $confidence(r) > confidence(\hat{r})$; so, the new measure keeps the accuracy of the widely-used measure confidence.

Measure $\tilde{\phi}$ 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). So, $\tilde{\phi}$ distinguishes positive association rules from the rest and also it is symmetry breaking. This new $\tilde{\phi}$ can create a completely useful test in just one step: $\mathbf{T}(r, \tilde{\phi}, \theta^*, \mathcal{D})$.

5.1. Evaluating the 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



$\tilde{\phi}$ against the old ϕ

Histogram of $\tilde{\phi}$

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

threshold θ^*, that should represent a perfect split between interesting rules and uninteresting rules (see proposition 3.1 and 3.2). So, now we study this convenenient value for θ^* for our measure $\tilde{\phi}$ (we know that this value exists by lemma 3.1).

The threshold θ^* only plays a role on the first part of the condition of the test, i.e, θ^* is just used to decide if the antecedent and consequent are correlated according to ϕ (P1). 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 for an input rule r: **H0) r is an uncorrelated association rule ($\tilde{\phi} = 0$)** 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^*.

For that, 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 realizacions 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 $\tilde{\phi}$

So, simulation of different samples will lead to a good approximation of the real distribution function, and we will be able to infer from threre 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 a good 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.

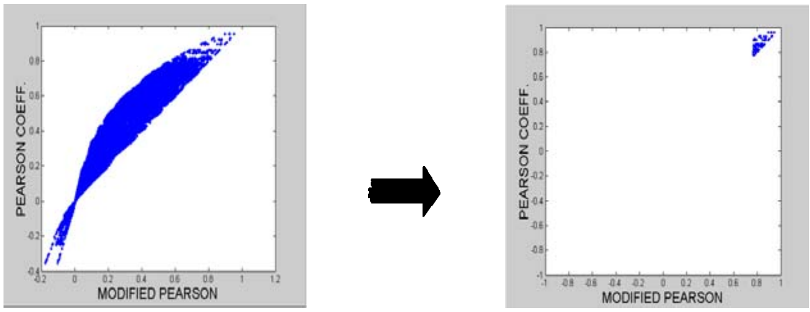


Figure 4: Process followed by the pruning strategy in real data

6. Experiments

Interestingly enough, our measure performs a completely useful test $\mathbf{T}(r, \tilde{\phi}, \theta^*, \mathcal{D})$, keeps the proper order among the surviving rules. This leads to the following one-step strategy.

Step 1/ 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. So, the strategy is not only simple, but also faster than any multi-test proposal. For synthetic data we generated synthetic 10,000 initial association rules such as in [16], 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[1] 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 the first graph of figure 4. The second graph of the same figure shows the 142 surviving rules after applying the proposed strategy. All these remaining rules have a confidence over 99%, so they are the strongest ones.

6.1. Overall Conclusion

In this paper we have presented a general framework that describes the last pruning phase of the uninteresting association rules. We formalize the optimal prune with a completely useful test created by a maximal harmless threshold, that is, a test formed by a measure IM capturing predicates P1 and P2 and keeping a proper order on rules. This formalization has allowed the evaluation of current different measures and the proposal of multi-test strategies. We also present a new measure, $\tilde{\phi}$, that meets all the necessary requirements for the optimal prune.

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 [15]). The proposals of this paper could also be followed in a temporal dimension (following the ideas in [12]).

References

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

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

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. "Fast Discovery of Association Rules". *Advances in Knowledge Discovery and Data Mining*, 307–328. 1996.

[3] Y.Bastide, N.Pasquier, R.Taouil, G.Stumme and L.Lakhal. "Mining Non-Redundant Association Rules using Frequent Closed Itemsets". *First International Conference on Computational Logic.* 2000.

[4] R.Bayardo Jr, and R.Agrawal. "Mining the Most Interesting Rules". *Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discvery and Data Mining*, 145–154. 1999.

[5] 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.

[6] S.Brin, R.Motwani, and C.Silverstein. "Beyond market baskets: Generalizing association rules to correlations". *Proc. ACM SIGMOD Int'l Conference on the Management of Data*, 1997.

[7] S. Brin, R. Motwani, J. Ullman, and S. Tsur. "Dynamic Itemset Counting and Implication Rules for Market Basket Data". *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, 1997.

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

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

[10] R.Hilderman, and H.Hamilton. "Evaluation of interestingness measures for ranking discovered knowledge". *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2001.

[11] B.Liu, W.Hsu, and Y.Ma. "Pruning and Summarizing the Discovered Associations". *ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining.* 1999.

[12] B.Liu, Y.Ma, and R.Lee. "Analyzing the Interestingness of Association Rules from the Temporal Dimension". *Proc. of the Int'l Conference on Data Mining.* 2001.

[13] S.Morishita, and J.Sese. "Traversing Itemset Lattice with Statistical Metric Pruning". *Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Database Systems*, 226–236. 2000.

[14] G.Piatetsky-Shapiro. "Discovery, analysis and presentation of strong rules". *Knowledge Discovery in Databases*, 229–248. 1991.

[15] A.Silberschatz, and A.Tuzhilin. "On Subjective Measures of Interestingness in Knowledge Discovery". *Knowledge Discovery in Databases.* 1995.

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

[17] P.Tan, V.Kumar, and J.Srivastava. "Selecting the Right Interestingness Measure for Association Patterns". *ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining.* 2002.

[18] M.Zaki. "Generating non-redundant Association Rules".In *Proc. of the Sixth Int'l Conference on Knowledge Discovery and Data Mining*, pages 34-43. 2000.

Appendix A

Pearson coefficient

The Pearson coefficient, ϕ, can be used to measure the degree of correlation between two variables (in our case antecedent and consequent of association rules). So, given the rule $A \Rightarrow C$, one can represent the following 2×2 contingency table as in table 3, where $A = 1$ represents the precence of the antecedent in transactions, and $A = 0$ its absence (equally for the consequent C). In fact, any association rule $A \Rightarrow C$ can be represented using the mentioned contingency table since each $\frac{f_{ij}}{|\mathcal{D}|}$, is $Pr(A = i \cup C = j, \mathcal{D})$.

Thus, for each association rule the degree of correlation between antecedent and consequent is:

	$C=1$	$C=0$	
$A=1$	f_{11}	f_{10}	f_{1+}
$A=0$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	DB

Table 3: Contingency table for two variables A and C

$$\phi(A,C) = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}}$$

Note that this metric is not symmetry breaking. When $\phi >> 0$ the correlation between A and C is highly positive (it is a strong positive rule); when $\phi < 0$ the correlation is negative (negative rule) and with $\phi \approx 0$ we have independence between the two variables (null association rule).

Appendix B

Here we provide definitions for Confidence [2], Conviction [7], Gini Index, Information Gain, J-Measure, Lift, PS [14], and IS [16], over a rule $r = A \Rightarrow C$. In order to make the definitions more readable, we will avoid the use of the $\mathcal{D}$ in each definition; so, we consider the $Pr(A)$ equivalent to the previous defined $Pr(A, \mathcal{D})$.

- $Confidence\ (A \Rightarrow C) = \frac{Pr(A \cup C)}{Pr(A)}$
- $Conviction\ (A \Rightarrow C) = \frac{Pr(A)Pr(\neg C)}{Pr(A \cup \neg C)}$
- The definition for the Gini Index is the following:

$$\begin{aligned} Gini\ (A \Rightarrow C) \quad &= \quad Pr(A)(Pr(C|A)^2 + Pr(\neg C|A)^2) + Pr(\neg A)(Pr(C|\neg A)^2 \\ &\quad\quad + Pr(\neg C|\neg A)^2) - Pr(C)^2 - Pr(\neg C)^2 \end{aligned}$$

- $Inf\ Gain\ (A \Rightarrow C) = \frac{H(A)+H(C)-H(A \cup C)}{H(A)}$
  where $H(A) = -Pr(A)\log Pr(A) - Pr(\neg A)\log Pr(\neg A)$, and $H(C) = -Pr(C)\log Pr(C) - Pr(\neg C)\log Pr(\neg C)$, and $H(A,C) = -\sum_i \sum_j Pr(A = i \cup C = j)\log Pr(A = i \cup C = j)$.
- The J-Measure of a rule $A \Rightarrow C$ is defined as:
  $J(A \Rightarrow C) = Pr(A)\Big(Pr(C|A)log\frac{Pr(C|A)}{Pr(C)} + Pr(\neg C|A)log\frac{Pr(\neg C|A)}{Pr(\neg C)}\Big)$
- $Lift\ (A \Rightarrow C) = \frac{Pr(A \cup C)}{Pr(A)Pr(C)}$
- $PS\ (A \Rightarrow C) = Pr(A \cup C) - Pr(A)Pr(C)$
- $IS\ (A \Rightarrow C) = \sqrt{Lift(A \Rightarrow C) \times Pr(A \cup C)}$

Appendix C

In order to make the proof more readable, we will avoid the use of the $\mathcal{D}$ in the definitions. We remind that a function $f(x)$ is said to be monotone in x if $x_1 < x_2$ implies that $f(x_1) \leq f(x_2)$.

Proof sketch of Lemma 3.2

To proof that these three conditions are sufficient for establishing a total order we must see that the following implication is always true:

$$r_1 <_1 r_2 \quad \text{or} \quad r_1 <_2 r_2 \quad \text{or} \quad r_1 <_3 r_2 \longrightarrow r_1 \leq_{IM} r_2$$

First of all, it is worth mentioning that the three partial orders we have previously defined ($r_1 <_1 r_2$, $r_1 <_2 r_2$, $r_1 <_3 r_2$) cannot occur at the same time over the same pair of rules. In other words, only a single of these partial orders (or, none of them) coexist over a given pair of rules r_1 and r_2. This observation is quite easy to justify following the definitions of the three partial orders; so, we are leaving the proof of this observation to the reader.

Now, suppose that we have a pair of rules $r_1 = A_1 \Rightarrow C_1$ and $r_2 = A_2 \Rightarrow C_2$, such that can be ordered with the first partial order, i.e, $r_1 <_1 r_2$. Then, consider a new rule $r = A \Rightarrow C$ where $Pr(A) = Pr(A_1) = Pr(A_2)$, $Pr(C) = Pr(C_1) = Pr(C_2)$, and $Pr(A_1 \cup C_1) < Pr(A \cup C)$ and $Pr(A \cup C) < Pr(A_2 \cup C_2)$.

Note that by definition $r_1 <_1 r$ and $r <_1 r_2$. Now, if a total order $\leq_{IM}$ has the first monotonicity property from the lemma 3.2, then $r_1 \leq_{IM} r$ and $r \leq_{IM} r_2$. Since total orders are transitive, we then have that $r_1 \leq_{IM} r_2$, which satisfies the claim.

We can use the same arguments in case the pair of rules are ordered by the second partial order, i.e, $r_1 <_2 r_2$. In this situation the second monotonicity property from lemma 3.2 is needed to proof the claim. Or also, in case that $r_1 <_3 r_2$, where we need $IM(r)$ to fulfill the third monotonicity property.