Alcom-FT Technical Report Series
ALCOMFT-TR-02-121

# A Multi-Step Pruning Strategy to Remove Uninteresting Association Rules *

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

May 24, 2002

**Abstract**

Association rule mining is a relevant task in Knowledge Discovery in Databases that aims at finding the group of items that most frequently appear with other groups of items in the database. Nowadays, many algorithmic strategies have been developed to discover all those association rules satisfying a minimum support constraint. However, the number of patterns extracted by these algorithms is often very large, leading to a rule quality problem, that is, out of the overwhelming number of discovered association rules only a small part of them are interesting.

This rule quality problem is currently addressed by specifying certain interestingness measures that are used as a single-step prune. The main problem of this solution is that not all measures are equally good at capturing the usefulness of a rule, and thus, after this pruning phase, some useless rules can still remain or also some useful rules can be removed permanently.

In this paper we investigate which properties the interestingness measures should keep in the last pruning phase. Later, we show how this prune turns out to be incomplete when only a single measure is chosen, and we introduce the proposal of a multiple-step pruning strategy that removes all useless rules and summarizes just the interesting ones. Specifically, we develop the working hypothesis that the Pearson coefficient is appropiate to guide the comparisons. The idea behind the strategy is keeping, at each pruning step, the good properties of the chosen measure, and avoiding, at the same time, its drawbacks. Finally, several experiments have been performed to show the viability and advantages of our method.

*This work is supported in part by EU ESPRIT IST-1999-14186 (ALCOM-FT), and matching funds TIC2000-1970-CE, and CIRIT 1997SGR-00366.

# 1. Introduction

One of the most relevant tasks in Knowledge Discovery in Databases is mining for association rules in large masses of data. This problem was first formulated by [2] and a very common application is market basket analysis, which aims at identifying the groups of items that are most frequently bought along with other groups of items. An example of an association rule in this market basket scope has the form *Cola* $\Rightarrow$ *Chips* [*Support* = 15%, *Confidence* = 80%], which means that 15% of the customers buy *Cola* and *Chips* together, and 80% of those who buy *Cola* buy *Chips* too.

When mining association rules of this form, a rule quality problem arises, that is, out of the great number of discovered association rules, only a small part of them are interesting. Currently, this problem is faced by calculating an interestingness measure over the rules: those rules ranking lower than a certain user-specified threshold on a certain measure will not be considered useful, and then, they will be pruned. The problem of this solution is that only a single measure is chosen, and not all measures are good at capturing the usefulness of a rule; so, the doubt is, which measure must be used to perform that sieve?

In this paper, we will discuss which properties the interestingness measures should keep in this last pruning phase. We will see how these measures do not suffice if they are used in a single-step prune and how most of them do not preserve the main studied properties. Finally, we investigate the proposal of a strategy that tries to summarize just the interesting patterns, regardless of the data domain and regardless of any other constraints on the initial support.

# 2. Problem Formulation

The "association rule" discovery task is often decomposed into two separate tasks:

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 first task is the most I/O intensive and up to the moment, a rich variety of algorithmic strategies have been proposed to find all the frequent itemsets: Apriori [3] and its variants [7] that traverse the induced lattice of itemsets in a breadth-first manner; or also algorithms traversing the lattice using a best-first strategy [4]; or finally, different approaches like FPGrowth [9] or TreeProjection [1] that try to reduce the database into a smaller structure.

The input of a frequent sets algorithm is a database composed of a collection of *transactions* or *itemsets*, each labeled by a unique identifier. In this database, 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 $Pr(I)$ be the ratio of the number of transactions in which $I$ appears to the number of all transactions in the database. We note the *support* of an itemset $I$ as $Pr(I)$. An itemset is called *frequent* if its support exceeds a given threshold, $\sigma$. The frequent sets problem can be stated as: given the database and the threshold $\sigma$, find *all* the frequent sets of items and their exact support. Logically, the lower the threshold $\sigma$, the greater number of discovered frequent itemsets.

In the second task, association rules are constructed from the maximal frequent sets discovered in the first phase, being maximal those frequent sets not included in any other one. 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 $X \cup Y = Z$. It is important to note that the support of the rule $X \Rightarrow Y$ is defined as the support of the itemset $X \cup Y$, that is, $Pr(X, Y)$.

## 2.1. Problems Concerning the Association Rule Discovery Task

As one can see, the number of generated rules can be quite large due to the big number of maximal discovered frequent sets and the subsequent combinatorial explosion to build all the possible rules. A feasible way of reducing that initial number of discovered frequent sets consists in setting the support threshold $\sigma$ with a high value in the first phase. The problem is that, with a high $\sigma$ the algorithm only discovers obvious frequent itemsets, and consequently, all the subsequent generated rules result to be not very novel. For this reason, in this paper we will avoid any suppositions on the initial threshold $\sigma$ used by the frequent sets algorithm, and we will assume that $\sigma$ can be set to any value. Actually, the worst case we will be facing is when $\sigma$ is fixed with a low value and many different rules are generated.

From these numerous constructed rules we know that only a very small fraction of the total are typically useful rules. Currently, association rules are selected or discarded according to some *interestingness measure*, that is, a criterion related to the supports of the constituting subsets of the antecedent and consequent of the rule. This means that the rule quality problem relies on the specification of a single interestingness measure being able to represent the significance of patterns. For instance, Agrawal et al [2] addressed the problem considering only those association rules that had *confidence* value no smaller than a user-specified threshold.

In fact, many other interestingness measures have appeared to replace the confidence in this post-pruning phase: conviction, Lift, Gini Index, Gain Ratio, J-Measure ... The main problem of any of them is that when they are used alone as a one-step prune, some useless rules can still hold or maybe some useful rules can be eliminated. These are "problematic rules" that must be carefully handled. It is known that limiting the initial value of $\sigma$ with an upper bound most of these mentioned "problematic rules" will be never generated since uncorrelated and negatively correlated itemsets are never discovered. However, we are interested in finding general pruning strategies with the ability to succeed on any domain and given any unconstrained $\sigma$.

### 2.2. When an Association Rule is Considered Interesting

In our problem, we are interested in rules $X \Rightarrow Y$ with antecedent and consequent positively correlated. Statistical measures such as the Pearson coefficient, that is $\phi$, or the $\chi^2$ test, are very useful to detect the correlation degree between two binary variables. So, the Pearson coefficient or the $\chi^2$ test could be used if we consider $X$ and $Y$ are binary variables representing the presence or the absence of antecedent/consequent of the rule in a transaction. However, it is worth to note that, in the market basket framework $\chi^2$ is not a reliable measure of association because it depends explicitly on the length of the database, so $\phi$ is used in later comparisons.

Thus, as a working hypothesis, we will consider an association rule *interesting* when it is a *positive association rule*, that is, its *association type is highly positive*. As mentioned below, this notion of *association type of a rule* $X \Rightarrow Y$, where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, has to do with the correlation degree between the antecedent, $X$, and the consequent, $Y$.

The correlation degree between antecedent and consequent can be measured via a $2 \times 2$ contingency table as in table 1, where $X = 1$ represents the precence of the antecedent in transactions, and $X = 0$ its absence (equally for the consequence $Y$). In fact, any association rule $X \Rightarrow Y$ can be represented using the mentioned contingency table since each $\frac{f_{ij}}{DB}$, being $DB$ the lenght of the database, is the support $Pr(X = i, Y = j)$.

| | $Y=1$ | $Y=0$ | |
|---|---|---|---|
| $X=1$ | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $X=0$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{+0}$ | $DB$ |

Table 1: Contingency table for the rule $X \Rightarrow Y$

Thus, for each association rule the degree of correlation between antecedent and consequent can be calculated, for example, with the Pearson coefficent (denoted with $\phi$).

$$\phi = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}}$$

When this correlation between $X$ and $Y$ is highly positive (that is $\phi >> 0$), then the rule $X \Rightarrow Y$ will be a *positive association rule*, i.e, an interesting rule. On the other hand, when the association is negative (i.e, $\phi < 0$) or close to zero (i.e, $\phi \approx 0$), then we consider that the rule should be pruned.

The reason to discard null association rules (i.e, uncorrelated association rules whose $\phi$ is close to zero) is because these imply independence between $X$ and $Y$, which makes the rule useless since we are looking for association patterns and not independent ones. On the other hand, the reason to discard negative association rules is that these rules should be better discovered with different specific algorithmic strategies, in particular those strategies having into account the negation of attributes, for example the bounded-neg-Apriori of [8], where the necessary monotonicity properties are preserved, which is not necessarily the case for statistical metrics [13].

## 3. Desired Properties of Measures in the Association Rule Framework

In order to study the behaviour of the main interestingness measures for association rules, it is important to introduce the general properties that they should satisfy. Considering all these properties we will later see how not all the current measures are equally good at performing a single-step prune.

It is important to note that these properties are to be applied over the association rules, not their original itemsets. In some other approaches like in [13] or [6], the main goal is to find the best correlated itemsets in the lattice. These itemsets suffice to represent the final discovered knowledge and no subsequent association rules are constructed. In [16] is presented a good general survey containing the main properties of interestingness measures for itemsets. In this paper, the only pruning measure applied over the itemsets is the support threshold $\sigma$, which is always necessary because it represents the statistical significance of the pattern. In this section, we restrict the properties of [16] into our association rule framework.

Let $IM$ be an interestingness measure for rules of the form $X \Rightarrow Y$ where $X \subset \mathcal{I}$ and $Y \subset \mathcal{I}$. The desired properties that $IM$ must accomplish in order to perform a perfect prune of all uninteresting rules in just one step are:

### (P1) $IM$ highly correlated with $\phi$

This property means that the interestingness measure $IM$ and the degree of correlation between the antecedent, $X$, and the consequent, $Y$, of the rule (represented by $\phi$ in the last section) must be highly correlated. The fulfilment of this property implies that $IM$ will be correlated with the type of association of the rule, and thus, $IM$ will be able to distinguish positive association from negative association of a given rule.

This property is quite important since negative association rules will be pruned only if we are able to find an $IM$ with the ability to distinguish them from the rest.

### (P2) $X \leftrightarrow Y$ permutation asymmetry

When it comes to association rules, there is a need to use asymmetric measures that can distinguish between the strength of implication for the rule $X \Rightarrow Y$ from $Y \Rightarrow X$, that is, a good $IM$ should not rank equally $X \Rightarrow Y$ and $Y \Rightarrow X$. More formally, the $X \leftrightarrow Y$ permutation assymetry of $IM$ implies that $IM$ is able to rank differently a given contingency table (representing a rule $X \Rightarrow Y$) from the transpose of that table (representing a rule $Y \Rightarrow X$).

Measures *not* satisfying this property should not be applied to association rules, only to their original itemsets.

(P3) $f_{11} \leftrightarrow f_{00}$ **and** $f_{01} \leftrightarrow f_{10}$ **permutation asymmetry**

This third property expresses how $IM$ should act 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, and vice-versa. Definitively, (P3) states that in the association rule framework, the co-presence of $X$ and $Y$ 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. Thus, association rules with greater value in the cell $f_{11}$ than in $f_{00}$, should be considered more interesting than those rules whose $f_{11} < f_{00}$.

The following three properties are very well-known and they were proposed by Piatetsky-Shapiro in [14].

(P4) $IM = 0$ **if** $X$ **and** $Y$ **are statistically independent**

(P5) $IM$ **monotonically increases with** $Pr(X,Y)$ **when** $Pr(X)$ **and** $Pr(Y)$ **are fixed**

(P6) $IM$ **monotonically decreases with** $Pr(X)$ **(or** $Pr(Y)$**) when the rest of the parameters (**$Pr(X,Y)$ **and** $Pr(Y)$ **or** $Pr(X)$**) are fixed**

Of these last three properties, (P4) has not to be compulsoryly satisfied in our framework. In fact, it suffices that $IM$ can distinguish uncorrelated association rules from the rest of rules in order to prune them. If $IM$ can do that, we do not care of which specific value is taken in case of that independence. For our purposes, we could even relax property (P4) in the following way: "$IM$ able to distinguish uncorrelated rules from correlated ones".

Table 2: **Properties satisfied by Measures for Association Rules**

| | **(P1)** | **(P2)** | **(P3)** | **(P4)** | **(P5)** | **(P6)** |
|---|---|---|---|---|---|---|
| $\phi$ | Yes | No | No | Yes | Yes | Yes |
| Confidence | No[1] | Yes | Yes | No | Yes | Yes |
| Conviction | No[1] | Yes | Yes | No | Yes | Yes |
| Gini Index | No[4] | Yes | No | Yes | No[4] | No[4] |
| Gain Ratio | No[4] | Yes | No | Yes | No[4] | Yes |
| J-Measure | No[4] | Yes | Yes | Yes | No[4] | No[4] |
| Lift | No[3] | No | Yes | No | Yes | Yes |
| $PS$ [14] | No[2] | No | No | Yes | Yes | Yes |
| $IS$ [15] | Yes | No | Yes | No | Yes | Yes |

[1]No, due to "false positives".

[2]No, due to "false negatives".

[3]No, due to "false positives" and "false negatives".

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

## 4. Behaviour of Current Interestingness Measures

This section introduces the behaviour of the main interestingness measures used to rank association patterns. A brief introduction to some of these measures can be found in [10], or in [11], and a good survey of them is also [15] and [16].

For comparison purposes, we generate artificial datasets such as in [15] containing $10,000$ random samples. Each sample is a $2\times 2$ contingency table representing an association rule $X \Rightarrow Y$ where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$ derived from a database of a given length $DB$. We can think of each table as a realization of the contingency table, as in table 1, for a rule $X \Rightarrow Y$ in the overall dataset. Each generated contingency table is subject, as in [15], to the following constraints: $f_{11}/DB < 1$, $f_{10}/DB < 1$, $f_{01}/DB < 1$ and $f_{11} + f_{10} + f_{01} < 1$. Apart from these restrictions, a minimum support $\sigma$ must be given and $f_{11} > \sigma$ in order to represent the support-based prune performed by the frequent sets algorithms of the first phase. In the following synthetic experiments we will assume $\sigma = 0$.

At a glance, comparisons of the main interestingness measures to $\phi$ can be grasped from figures 1 and 2. Note that the rules we want to keep are exactly those in the high top half of each square, that is, with $\phi >> 0$.

### 4.1. Confidence

The *confidence* of a rule $X \Rightarrow Y$ is defined as:

$$Confidence\ (X \Rightarrow Y) = \frac{Pr(X,Y)}{Pr(X)}$$

It can take values from $[0,1]$ and its definition is in fact the conditional probability of $Pr(Y|X)$. It measures strenght of the association rule.

As shown in [7] the confidence has some important problems, in particular, it ignores $Pr(Y)$. For instance, $\frac{Pr(X,Y)}{Pr(X)}$ could equal $Pr(Y)$, i.e, the occurrence of consequent $Y$ is unrelated to the antecedent $X$, which forms a null association rule. This value could still be high enough to make this uninteresting rule hold.

Two consequences of this problem are worth mentioning. First, confidence is not always good at pruning those problematic rules called "false positives" (i.e, null association rules with high value of confidence). "False positives" should be removed, but confidence-based prune can still keep them. Second, neither property (P1) nor property (P4) are satisfied by this measure.

Confidence has been a widely used measure in this association rule framework. In fact, in [5] it was proved that the best association rule according to any interestingness measure must reside along the support/confidence border.

### 4.2. Conviction

The *conviction* of a rule $X \Rightarrow Y$ was proposed in [7] and can be defined as:

$$Conviction\ (X \Rightarrow Y) = \frac{Pr(X)Pr(\neg Y)}{Pr(X,\neg Y)}$$

Conviction can take values in the range $[0, +\infty)$, being 1 the value implying independence between antecedent and consequence.

As it can be seen in the first graph of figure 1, conviction suffers from "false positives" (i.e, null association rules with a high value of conviction). These "false positive" rules run the risk of not being removed even with some conviction threshold over 1. Once again, conviction is a risky measure when used as a single-step prune.

An advantage of this measure is that those rules with conviction under 1 can be always fairly pruned since all of them are negative association rules (see first graph of figure 1). Another advantage is that it assigns a value of $+\infty$ for perfect implications and its wide range of values makes conviction the most discriminatory measure.

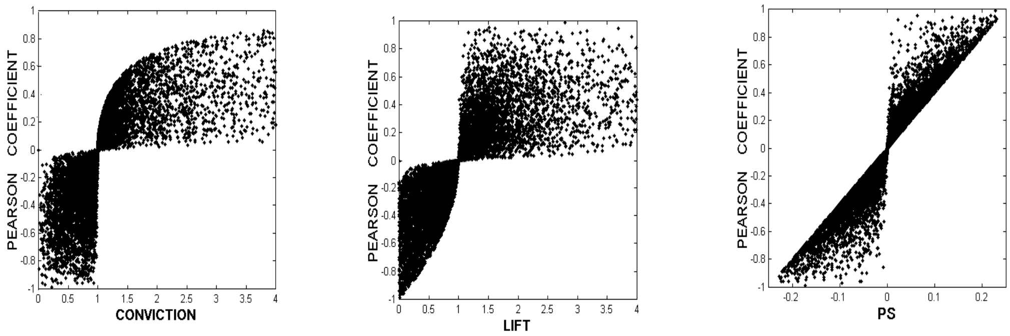


Figure 1: Correlation of Conviction, Lift and $PS$ against $\phi$

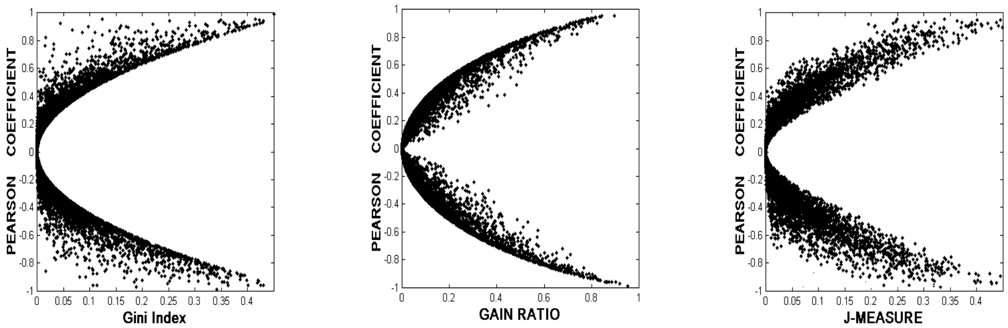


Figure 2: Correlation of Gini Index, Gain Ratio and J-Measure against $\phi$

### 4.3. Gini Index

The *Gini Index* of a rule $X \Rightarrow Y$ is given by the expression:

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

This value may range from 0 (when antecedent and consequent are completely uncorrelated) to 0.5 (for perfect correlation).

In first graph of figure 2 it is shown the relationship between the Pearson coefficient and Gini Index regarding the synthetic sample of rules. As one can observe, the relationship is symmetric in the horizontal axis, and we can conclude that Gini Index does not distinguish between positive and negative association rules. This fact makes the Gini Index a measure unsatisfying the properties (P1),(P5) and (P6) as can be seen in table 2.

### 4.4. Information-based Measures

This type of interesetingness measures are born from information theory. In particular, we are considering for our purposes the *Gain Ratio* and the *J-Measure*.

The Gain Ratio of a rule $X \Rightarrow Y$ can be defined as:

$$Gain\ Ratio\ (X \Rightarrow Y) = \frac{H(X) + H(Y) - H(X,Y)}{H(X)}$$

where $H(X) = -Pr(X)\log Pr(X) - Pr(\neg X)\log Pr(\neg X)$, and $H(Y) = -Pr(Y)\log Pr(Y) - Pr(\neg Y)\log Pr(\neg Y)$, and $H(X,Y) = -\sum_i \sum_j Pr(X=i, Y=j)\log Pr(X=i, Y=j)$.

The J-Measure of a rule $X \Rightarrow Y$ is defined as:

$$J(X \Rightarrow Y) = Pr(X)\Big(Pr(Y|X)log\frac{Pr(Y|X)}{Pr(Y)} + Pr(\neg Y|X)log\frac{Pr(\neg Y|X)}{Pr(\neg Y)}\Big)$$

Gain Ratio ranges between 0 (for independence between antecedent and consequent) to 1 (for perfect association). The range of J-Measure is between 0 (for independence) and 0.5 (for perfect association).

Graphs second and third of figure 2 show the relationship between the Pearson coefficient and the Gain Ratio, and the same for J-Measure. As one can observe, both measures treat both positive and negative association rules in the same way. Again, we are in front of two measures that do not distinguish the type of association of a rule.

Since we are only interested in highly positive association rules, a measure treating both positive and negative association in the same way is completely useless in a single-step prune. However, it is important to note that in case one could previously eliminate all negative association rules, then (P1),(P5) and (P6) would be satisfied by Gini index, Gain Ratio and J-Measure.

### 4.5. Lift

The interestingness measure *Lift* of a rule $X \Rightarrow Y$, can be defined in terms of the lift of the itemsets of the rule. That is,

$$Lift\ (X \Rightarrow Y) = \frac{lift(X \cup Y)}{lift(X)lift(Y)}$$

where the definition of lift for a set is $lift(x_1, x_2, \ldots x_n) = \frac{Pr(x_1, x_2, \ldots x_n)}{Pr(x_1)\ldots Pr(x_n)}$.

This definition of Lift implies that the best rule that can be formed from a given itemset is the one maximizing the ratio between $lift(X \cup Y)$ and $lift(X)lift(Y)$. Then, this measure can result appropriate to sort those interesting rules coming from the same original itemset.

Lift for association rules ranges between $[0, +\infty)$, being 1 the value indicating the independence between the antecedent and the consequent. Values greater that 1 imply "non-negative" association rules.

The second graph of figure 1 shows the relationship between Lift and the Pearson coefficient. It can be seen that, in case Lift was used in a single-step prune, it could suffer from keeping "false positives" (i.e, null association rules with a high value of Lift) and removing "false negatives"(i.e, positive association rules with a value of Lift close to 1). However, it is important to note that a more important problem of Lift is the non-fulfilment of property (P2).

### 4.6. Piatetsky-Shapiro Measure

This measure was introduced by Piatetsky-Shapiro in [14] as an example of a measure that held the properties (P4), (P5) and (P6). The definition of this interestingness measure is the following:

$$PS\ (X \Rightarrow Y) = Pr(X,Y) - Pr(X)Pr(Y)$$

This measure ranges between $[-0.25, 0.25]$. When the antecedent $X$ and the consequent $Y$ of the rule are independent $PS$ equals 0. On the other hand, positive association rules, thus interesting rules, have a value of $PS$ greater than 0.

As figure 1 shows, one problem of the Piatetsky-Shapiro measure are the "false negative" rules (i.e, positive association rules with a value of $PS$ close to 0). This can cause the removal of interesting rules when one is fixing a high $PS$ threshold. A more difficult problem to solve is the non-fulfilment of (P2) and (P3), which makes $PS$ less applicable to the association rule framework.

### 4.7. $IS$ Measure of Tan-Kumar

This useful interestingness measure was proposed in [15] and it derives from the relation of Lift with the Pearson coeficient considering certain restrictions of the initial support $\sigma$.

The definition of the $IS$ measure of [15] is the following:

$$IS\ (X \Rightarrow Y) = \sqrt{Lift(X \Rightarrow Y) \times Pr(X,Y)}$$

The values of $IS$ can range from $[0,1]$. The value for the complete independence between the antecedent and the consequent of the rule is when $IS = \sqrt{Pr(X,Y)}$. This last fact violates the property (P4) since the value of $Pr(X,Y)$ depends on each particular rule.

$IS$ has the advantage of being highly correlated with the Pearson coefficient, which is very important in order to distinguish positive from negative association of patterns. The main disadvantage of using $IS$ for association rules is the non-fulfilment of property (P2).

## 5. Conclusions

From the above presented measures one can see how *none* of them is able to perform a perfect prune of all uninteresting rules at a time. The reasons are:

1. some measures have problems with "false negatives" (i.e, $PS$ and Lift) which make them very risky since one can even remove positive association rules;
2. others have problems with "false positives" (i.e, conviction, confidence and Lift) and they could maintain null association rules after the pruning step;
3. finally, other measures fail in the satisfaction of certain important properties (for example, Pearson coefficient, Lift, $IS$, $PS$, Gain ratio or J-Measure). For instance, Pearson coefficient could be the perfect interestingness measure for association rules in a single-step prune as long as it would satisfy properties (P2) and (P3). However, this measure is directly discarted in our framework due to this failure.

These three main sorts of drawbacks of measures make the widely-used single-step prune incomplete, and motivates the investigation of a possible multiple-step pruning strategy. The idea of this strategy is performing at each step an action that can ease the subsequent prune of the next step. For example, we could use conviction in one of the steps of the strategy as long as all "false positive" rules were previously removed, or also, we could use J-Measure in another step as long as negative association rules were previously eliminated.

It is important to note that some of the above introduced measures will be never chosen in our pruning strategy. These rejected measures are Pearson coefficient, Gain ratio, Gini Index, $IS$ and $PS$. The main reason is that they do not satisfy properties (P2) and/or (P3), which are a must given our assymmetric framework. Note that measure Lift is not included in this list of rejected measures: although Lift does not fulfil (P2) it will be used in a special situation that does not affect the rest of the strategy, as we will later see.

## 6. Our Pruning Strategy Proposal

The idea of our pruning strategy proposal is using a different measure at each pruning step, but of course, avoiding its drawbacks. So, in order to accomplish this goal, at each step we perform the elimination of association rules that not only are considered uninteresting, but also can be a problem for the measure chosen in the next pruning step.

Given our assymmetric framework we want properties (P2),(P3),(P5) and (P6) hold during the whole multiple-step strategy, that is, we want that every chosen measure in the strategy fulfil these properties at each pruning step. On the other hand, properties (P1) and (P4) do not need to be satisfied by every measure at each step. We know that their fulfiment was a must in case of the single-step prune because there was the need to separate interesting rules from the rest at a time, using just one measure. However, in our proposal this separation goal will be achieved through the different steps that our strategy will perform, not in just one single prune. Thus, the strategy ensures that at the end of the last pruning step only interesting rules will hold, making (P1) and (P4) an unnecessary properties.

In the following proposed strategy, each step represents an elimation phase of a set rules. Each elimination phase is performed over the resulting set of rules from the preceeding step.

**0/** Get an initial set of rules containing all those rules generated from itemsets whose support is over a user specified $\sigma$.

**1/** Fix a certain confidence threshold, for instance, prune all those rules whose $confidence < 0.5$.

Although confidence is a mesure that does not fulfil properties (P4) nor (P1), this first pruning step can be justified following the ideas in [5]. In particular, [5] proved that the best association rules according to any interestingness measure reside along the support/confidence border. Then, the use of confidence as our first pruning step guarantees that the hidden interesting rules will reside inside this set of remaining unpruned rules.

However, as we have seen in the previous section, confidence does not ensure a good prune, and along with the hidden interesting rules can still reside also some negative association rules or null association rules with a high confidence. These rules are not eliminated in this phase but they deserve to be pruned in subsequent steps.

It is worth to note that the confidence threshold used in this step is mainly user-dependent. Nevertheless, regarless of the value for this user-specified threshold, the strategy is always equally valid. In fact, we propose here a very low threshold (i.e, 0.5) since, as we will later see, the whole strategy will ensure that all the final unpruned rules are also the strongest ones in terms of confidence.

**2/** Prune all negative association rules, that is, remove rules whose $conviction < 1$.

As we are not interested in negative association rules one can prune them by eliminating all those rules whose conviction is under the value of 1. Once this is done, remaining rules have convicition greater than 1, which means that, up to this point, all the remaining unpruned rules are positive association rules or also null association rules.

It is important to note that the goal of this step is removing *only* negative association rules, which can be clearly achieved fixing 1 as threshold. Although a threshold greater than 1 would also guarantee the elimination of all negative association rules, we still would keep on holding null association rules. So, the removal of these uncorrelated rules will be left for the next steps and this second step only focuses on the removal of negative association rules.

Regarding the necessary properties, conviction always satisfies (P2) to (P6) (except (P4) which is not so important in this point if the strategy can have removed all null association rules at the end of the last step).

**3/** Prune uncorrelated association rules, that is, remove all rules whose *J-Measure* $< 0.2$.

Since step 2 of the strategy can guarantee that negative association rules are not present up to that point, one can use measures like J-Measure, Gini Index or Gain Ratio (measures that treated equally negative and positive association rules). The reason is that if negative association rules are not present the dicotomy disappears.

Apart from that, when negative association rules are not present the important properties (P1) and (P4) are *always* satisfied by these measures (see table 2). The satisfaction of these properties can allow us to eliminate all null association rules from the set. It is worth to note that this step could not have been performed if negative association rules had not been eliminated in the previous step.

Gini Index and Gain Ratio are rejected for our purposes due to the non-fulfilment of (P3), which is a compulsory property given the framework. But on the other hand, note that when negative association rules are eliminated J-Measure satisfies all the properties.

The threshold of the pruning step, *J-Measure*$<0.2$, is a parameter selected from the synthetic experiments (see the third graph of figure 2). In particular, the graph shows that from this point on all positive association rules are highly correlated.

### 4/ Prune and Sort by conviction

Up to this point one can be sure that negative and null association rules have been removed from the original set of rules. So the remaining set is certainly formed only by positive association rules. Since no null association rules are present, then "false positives" have disappeared too. So, if up to this point still many rules remain, it is the perfect step to use conviction as the last pruning measure.

Apart from that, since conviction is the most discriminatory measure, it is very useful to sort the remaining rules of the last prune by conviction. This sorting operation provides a useful ranking of rules where the most interesting ones are those with a high conviction value.

### 5/ In case rules coming from the same itemset would have the same conviction values, sort them by Lift

This last step of the strategy includes those cases where rules coming from the same original itemset have the same value of conviction, that is, they will rank equally in the step 4. A possible solution is using the measure of Lift to undo the draw, since, by its own definition, Lift turns out to be very appropriate to distinguish those rules constructed from the same itemset. So, sorting them by Lift one can choose the best of them when conviction does not.

The fact that Lift does not fulfil (P2) is not a drawback in this last point. The reason is that, with joined steps 4 and 5 of the strategy, one is performing a sorting operation by the key *conviction + Lift*: i.e, it is conviction the measure mantaining the main order, and Lift just is used in case of a draw. So, the assymetry of (P2) of rules is preserved by the first sort performed in the step 4.

## 7. Experiments

With the aim of studying the behaviour of our proposal, experiments using synthetic samples and also real data have been performed. To analyze the effectiveness of the pruning strategy we look at the number of removed association rules after each step and the correlation that the remaining association rules have with the Pearson coefficient.

For synthetic data we generated 10,000 initial association rules such as in [15], considering that the minimum support threshold is $\sigma = 0$. In figure 3 we can follow graphically the process of the pruning strategy after each step. After applying the pruning step one, 3,251 rules still hold. The first graph in figure 3 we can see the Pearson coefficient of the remaining rules against conviction.

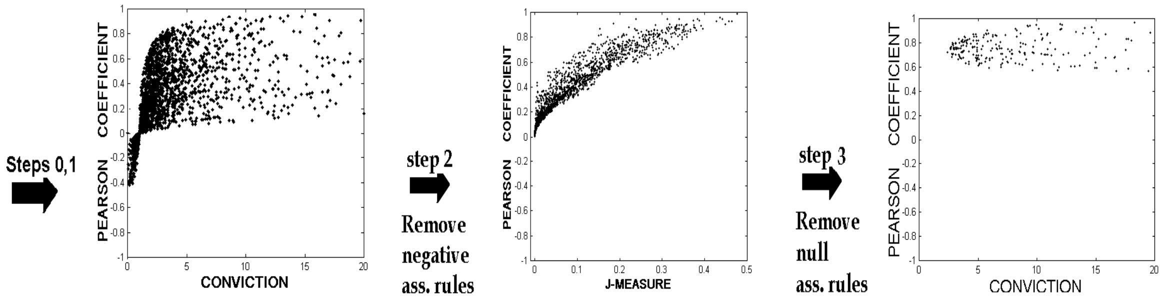


Figure 3: Process followed by the pruning strategy in synthetic data

We can observe how, obviously, null and negative association rules still remain (moreover, some null association rules are "false positives").

Step 2 removes all negative association rules. However, still 1,644 rules remain. In second graph of figure 3 we plotted these remaining rules against J-Measure. We observe how clearly negative association rules do not exist any longer, which eases the use of J-Measure as separator of null association rules from the rest.

With step 3 the strategy removes the uncorrelated association rules, and only 250 rules hold. In last graph of figure 3 these remaining rules are plotted against conviction. Observe that "false positives" have disappeared, which eases the use of conviction as the last sorting measure. In fact, if many rules still remain at this point, one can keep on removing them by incrementing the minimum threshold of conviction.

Performing sorting steps 4 and 5 one can rank the surviving rules. In this synthetic experiment, the first 50 rules have conviction greater than 20 and a confidence greater than 98%. So, the final conclusion of this experiment is that we have got *only* the strongest positive association rules.

The next goal was 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. The results of the strategy can be followed in figure 4.

In contrast with synthetic experiments, we used a $\sigma = 0.15$ and we got almost no negative association rules (as can be seen in the first and second graph of figure 4). However, the number of remaining rules after steps 0,1 and 2 are still 50,927, that is, a great important number. Step 3 is in charge of removing null association rules, and then we get 4,260 positive association rules. Graphically, these last remaining rules are represented in third graph of figure 4. It is worth to note that the third graph of figure 4 is equivalent to first graph of the same figure but without negative nor null association rules. This fact implies that the pruning strategy has fairly removed all the uninteresting rules avoiding the "false positives" and the "false negatives" problems.

After step 3, we can still perform another prune using conviction, since "false positives" have completely gone in previous steps. In case we perform a sorting operation by conviction over the 4,260 rules, then we can choose the most important ones. For instance, the 50 first rules after performing step 4 and 5 have a confidence over 99%, so they are the strongest rules.

# References

[1] R. Agarwal, C. Aggarwal, and V. Prasad. "A Tree Projection Algorithm for Generation of Frequent Item Sets". *Journal of Parallel and Distributed Computing*, 350–371. 2000.

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

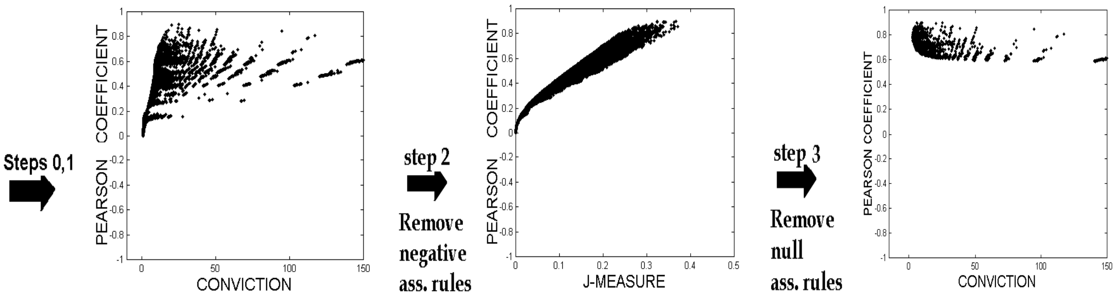


Figure 4: Process followed by the pruning strategy in real data

[2] 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.

[3] 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.

[4] J.Baixeries, G.Casas-Garriga, and J.L.Balcázar. "A Best First Strategy for Finding Frequent Sets". *Extraction et gestion des connaissances* (EGC'2002), 100–106. 2002.

[5] 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.

[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*, 255–264. ACM Press, 1997.

[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] J. Han, J. Pei, and Y. Yin. "Mining Frequent Patterns Without Candidate Generation". *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, 2000.

[10] R.Hilderman, and H.Hamilton. "Knowledge Discovery and Interestingness Measures: a Survey". Technical Report TR-99-04, Dept. of Computer Science, University of Regina, 1999.

[11] 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*, 247–259, 2001.

[12] 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.

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

[16] 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.