Closed-set-based Discovery of Bases of Association Rules

José L Balcázar, Diego García-Saiz, Domingo Gómez-Pérez and Cristina Tîrnăucă

Abstract The output of an association rule miner is often huge in practice. This is why several concise lossless representations have been proposed, such as the "essential" or "representative" rules. We revisit the algorithm given by Kryszkiewicz (Int. Symp. Intelligent Data Analysis 2001, Springer-Verlag LNCS 2189, 350–359) for mining representative rules. We show that its output is sometimes incomplete, due to an oversight in its mathematical validation. We propose alternative complete generators and we extend the approach to an existing closure-aware basis similar to, and often smaller than, the representative rules, namely the basis $\mathscr{B}^*_{\tau,\gamma}$.

1 Introduction

Association rule mining is among the most popular conceptual tools in the field of Data Mining. We are interested in the process of discovering and representing regularities between sets of items in large scale transactional data. Syntactically, the association rule representation has the form of an implication, $X \to Y$; however, whereas in Logic such an expression is true if and only if Y holds whenever X does,

José L Balcázar
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain, e-mail: joseluis.blacazar@unican.es

Diego García-Saiz
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain, e-mail: diego.garciasuc@alumnos.unican.es

Domingo Gómez-Pérez
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain, e-mail: domingo.gomez@unican.es

Cristina Tîrnăucă
Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain, e-mail: cristina.tirnauca@unican.es

an association rule is a partial implication, in the sense that it is enough if Y holds *most of the times* X does.

To endow association rules with a definite semantics, we need to make precise how this intuition of "most of the times" is formalized. There are many proposals for this formalization. One of the frequently used measures of intensity of this kind of partial implication is its *confidence*: the ratio between the number of transactions in which X and Y are seen together and the number of transactions that contain X. In most application cases, the search space is additionally restricted to association rules that meet a minimal *support* criterion, thus avoiding the generation of rules from items that appear very seldom together in the dataset (formal definitions of support and confidence are given in Section 2.1).

Many association rule miners exists, Apriori (see [Agrawal et al., 1996]) being one of the most widely discussed and used. The major problem shared by all mining algorithms is that, in practice, even for reasonable support and confidence thresholds, the output is often huge. Therefore, several concise lossless representations of the whole set of association rules have been proposed. These representations are based on different notions of "redundancy". In one of these, a rule is redundant if it is possible to compute exactly its confidence and support from other information such as the confidences and supports of other *informative* rules (see [Kryszkiewicz, 2002, Luxenburger, 1991, Hamrouni et al., 2008, Pasquier et al., 2005]); this is a quite demanding property. We settle for a weaker version proposed in several works; informally, in that version, a rule is *redundant* with respect to another one if its confidence and support are always greater, in *any* dataset. To avoid this redundancy, exactly one notion has been identified in several sources, namely the *representative rules*; and a closure-aware variant both of the redundancy notion and of the redundancy-free basis is given in [Balcázar, 2010a] (precise definitions and references are given below).

We focus in this paper on the main results of [Kryszkiewicz, 2001], where a purportedly faster algorithm to construct representative rules is given, and show by an example that that algorithm is not guaranteed to always output all representative rules, because it is based on a property that does not hold in general; namely, the characterization of the frequent closed sets that admit a decomposition into representative rules misses some such sets. We propose an alternative, complete characterization, leading us to the proposal of a first alternative algorithm that is guaranteed to output all the representative rules: we pre-compute, for each closed set, some parameters that depend on the confidence and support thresholds, and then use the above mentioned new characterization to generate all representative rules. Compared to the potentially incomplete algorithm in [Kryszkiewicz, 2001], this algorithm, guaranteed to be complete, has a main drawback: in [Kryszkiewicz, 2001], the internal local parameters only depend on the support threshold, but in our algorithm these parameters depend also on confidence. Therefore, each time a new confidence threshold is introduced by the user, the algorithm has to redo all computations. Thus, we provide a second algorithm, composed of two parts: the first one is a pre-processing phase, dependent only on support, in which a subdivision of the interval $(0,1]$ is associated to each closed itemset, and the second part uses

this partition to determine, for a given value of the confidence threshold, which are those sets that can generate representative rules.

Then, we extend the process to a similar basis which profits from the more powerful redundancy notions available for full-confidence implications to often obtain smaller bases in many applications.

There are a couple of subtle differences between one of the usual definitions of association rule (the one we employ) and the one in [Kryszkiewicz, 2001]. First, we do allow having rules with empty antecedent (clearly, all of them have confidence equal to the normalized support of the consequent). Moreover, we do not require the inequalities to be strict when imposing a given support and confidence threshold. This is just a small detail that comes handy when the user is interested in obtaining the set of all representative rules of confidence 1. However, we have carefully tuned all our argumentations in such a way that these differences are not relevant; for instance, we have chosen a counterexample that invalidates Property 9 of [Kryszkiewicz, 2001] independently of which of the two definitions is used.

The article is structured as follows. In Section 2 we introduce the basic notions and notations that will be used throughout the paper and part of the contents of [Kryszkiewicz, 2001]; and we show that the algorithm provided there is not guaranteed to always provide the whole set of representative rules. In Section 3 we define new parameters and discuss their usefulness in generating the set of all representative rules, providing also efficient algorithms for this task. We describe in Section 4 a parallel development for an alternative basis, often smaller than the representative rules. Section 5 contains a comparison of our approach with the one in [Kryszkiewicz, 2001] on some datasets. Concluding remarks and further research topics are presented in Section 6.

2 Preliminaries

A given set of available items $\mathscr{U}$ is assumed; subsets of it are called itemsets. We will denote itemsets by capital letters from the end of the alphabet, and use juxtaposition to denote union, as in XY. The inclusion sign as in $X \subset Y$ denotes proper subset, whereas improper inclusion is denoted $X \subseteq Y$. For a given dataset $\mathscr{D}$, consisting of n transactions, each of which is an itemset labeled with a unique transaction identifier, we define the *support* $sup(X)$ of an itemset X as the ratio between the cardinality of the set of transactions that contain X and the total number of transactions n. An itemset X is called *frequent* if its support is greater than or equal to some user-defined threshold $\tau \in (0,1]$. We denote by $F_\tau = \{X \subseteq \mathscr{U} \mid sup(X) \geq \tau\}$ the set of all frequent itemsets.

Given a set $X \subseteq \mathscr{U}$, the *closure* $\overline{X}$ *of* X is the maximal set (with respect to the set inclusion) $Y \subseteq \mathscr{U}$ such that $X \subseteq Y$ and $sup(X) = sup(Y)$. It is easy to see that $\overline{X}$ is uniquely defined. We say that a set $X \subseteq \mathscr{U}$ is *closed* if $\overline{X} = X$.

Closure operators are characterized by the three properties of extensivity: $X \subseteq \overline{X}$; idempotency $\overline{\overline{X}} = \overline{X}$; and monotonicity: $\overline{X} \subseteq \overline{Y}$ if $X \subseteq Y$. Moreover, intersections of

closed sets are closed. The empty set is closed if and only if no item appears in each and every transaction.

A *minimal generator* is a set X for which all proper subsets have closures different from the closure of X (equivalently, X is a minimal generator if and only if $sup(Y) > sup(X)$ for all $Y \subset X$).

Also, $FC_\tau = \{X \in F_\tau \mid \overline{X} = X\}$ represents the set of all frequent closed sets, and $FG_\tau = \{X \in F_\tau \mid \forall Y \subset X, sup(Y) > sup(X)\}$ is the set of all frequent minimal generators. Note that FC_τ constitutes a concise lossless representation of frequent itemsets, since knowing the support of all sets in FC_τ is enough to retrieve the support of all sets in F_τ.

Example 1. Let $\mathscr{D}$ be the dataset represented in Table 1 where the universe $\mathscr{U}$ of attributes is $\{a,b,c,d,e,f\}$, and consider the threshold $\tau = 0.15$. Clearly, all subsets of $\mathscr{U}$ are frequent, $FC_\tau = \{\emptyset, a, b, c, ab, ac, ad, bc, abcde, abcdef\}$ and $FG_\tau = \{\emptyset, a, b, c, d, e, f, ab, ac, bc, bd, cd, abc\}$ (we abuse the notation and denote sets by the juxtaposition of their constituent elements).

Table 1 Dataset $\mathscr{D}$

a	b	c	d	e	f
1	1	1	1	1	1
1	1	1	1	1	0
1	1	0	0	0	0
1	0	1	0	0	0
0	1	1	0	0	0
1	0	0	1	0	0

2.1 Association Rules and Representative Rules

Given X in F_τ, the following two notions were introduced in [Kryszkiewicz, 2001] (with longer names):

$$mxs_\tau(X) = \max(\{sup(Z) \mid Z \in FC_\tau, Z \supset X\} \cup \{0\}),$$
$$mns_\tau(X) = \min(\{sup(Y) \mid Y \in FG_\tau, Y \subset X\} \cup \{\infty\}).$$

That is, $mxs_\tau(X)$ represents the maximum support of all proper frequent closed supersets of X, and $mns_\tau(X)$ is the minimum support of minimal generators that are proper subsets of X. The extra 0 and ∞ are added in order to make sure that $mxs_\tau(X)$ and $mns_\tau(X)$ are defined even for the cases in which X has no proper supersets that are frequent and closed, or when it does not have proper subsets that are minimal generators. It is easy to check that $mxs_\tau(X) \leq sup(X) \leq mns_\tau(X)$. Moreover, in [Kryszkiewicz, 2001] it is shown that:

Proposition 1. *Given $\tau \in (0,1]$ and an itemset $X \in F_\tau$, X is closed if and only if $sup(X) > mxs_\tau(X)$ and X is a minimal generator if and only if $sup(X) < mns_\tau(X)$.*

The association rules considered in this work are implications of the form $X \to Y$, where $X, Y \subseteq \mathscr{U}$, $Y \neq \emptyset$ and $X \cap Y = \emptyset$. In [Kryszkiewicz, 2001], rules with $X = \emptyset$ are disallowed, but we do permit them as in practice such rules often play a useful role related to coverings, described below. The *confidence* of $X \to Y$ is $conf(X \to Y) = sup(XY)/sup(X)$, and its *support* is $sup(X \to Y) = sup(XY)$. The problem of mining association rules consists in generating all rules that meet the minimum support and confidence threshold criteria, i. e. enumerate the following set: $AR_{\tau,\gamma} = \{X \to Y \mid sup(X \to Y) \geq \tau, conf(X \to Y) \geq \gamma\}$.

Since the whole set of association rules is quite big in real-world applications, a number of formalizations of the notion of *redundancy* among association rules have been introduced (see [Aggarwal and Yu, 2001, Balcázar, 2010a, Kryszkiewicz, 1998b, Pasquier et al., 2005, Phan-Luong, 2001, Luxenburger, 1991, Zaki, 2004, Cristofor and Simovici, 2002], the survey [Kryszkiewicz, 2002], and Section 6 of [Ceglar and Roddick, 2006]). In one common approach, the *cover* set $C(X \to Y)$ of a rule $X \to Y$ is defined by $C(X \to Y) = \{X' \to Y' \mid X \subseteq X' \text{ and } X'Y' \subseteq XY\}$. Such rules $X' \to Y'$ are redundant with respect to $X \to Y$ in the following sense (see [Aggarwal and Yu, 2001, Kryszkiewicz, 1998b] and also [Kryszkiewicz, 1998a, Balcázar, 2010a, Phan-Luong, 2001]):

Proposition 2. *Let r, r' be association rules. Then $r' \in C(r)$ implies $sup(r') \geq sup(r)$ and $conf(r') \geq conf(r)$.*

In fact, this implication is a full characterization, that is, if r' has always at least the same confidence and at least the same support as r then it must belong to the cover set. Avoiding such redundancies leads to the set $RR_{\tau,\gamma}$ of *representative association rules*. A rule r in $AR_{\tau,\gamma}$ is said to be *representative*, or *essential*, if it is not contained in the cover set of any other rule in $AR_{\tau,\gamma}$, i. e.

$$RR_{\tau,\gamma} = \{r \in AR_{\tau,\gamma} \mid \forall r' \in AR_{\tau,\gamma} (r \in C(r') \Rightarrow r = r')\}.$$

Proposition 3. *The following properties hold:*

- $RR_{\tau,\gamma} = \{X \to Y \in AR_{\tau,\gamma} \mid \neg\exists X' \to Y' \in AR_{\tau,\gamma}, (X = X', XY \subset X'Y') \text{ or } (X' \subset X, XY = X'Y')\}$
- *if $X \to Z \backslash X$ with $X \subset Z$ is in $RR_{\tau,\gamma}$ then $Z \in FC_\tau$ and $X \in FG_\tau$.*

Therefore, any algorithm that aims at the discovery of all representative rules should consider only rules of the form $X \to Z \backslash X$ with $X \subset Z$, $Z \in FC_\tau$ and $X \in FG_\tau$. Clearly, not all sets in FC_τ can be decomposed in such a way, and one should look only into those that do.

Example 2. Consider the dataset in Example 1. The set ad is both frequent and closed, but none of the rules $a \to d$, $d \to a$ or $\emptyset \to ad$ are representative given the thresholds $\tau = 0.15$ and $\gamma = 0.33$: $a \to d$ is in the cover set of $a \to bd$, $d \to a$ is in

the cover set of $d \to ab$ and $\emptyset \to ad$ is in the cover set of $\emptyset \to abd$. Also, it is easy to check that, at $\tau = 0.15$ and $\gamma = 0.4$, one can obtain representative rules exactly out of the following closed sets: ab, ac, ad, bc, $abcde$, and $abcdef$.

So, if we denote by $RI_{\tau,\gamma}$ the set of all frequent closed itemsets from which at least one representative rule can be generated, one possible approach to representative rule mining is to synthesize first the set $RI_{\tau,\gamma}$, and then, for each element Z in $RI_{\tau,\gamma}$, to find non-empty subsets X such that $X \to Z \backslash X$ is representative. This is precisely the idea behind Algorithm *GenRR* in [Kryszkiewicz, 2001]. The problem there is that the characterization of the set $RI_{\tau,\gamma}$ given by Property 9 of the same paper (on page 355) is incorrect, possibly leaving out some of the sets that can lead to representative rules. Namely, it is stated that $RI_{\tau,\gamma} = \{X \in FC_\tau \mid sup(X) \geq \gamma * mns_\tau(X) > mxs_\tau(X)\}$; right-to-left inclusion indeed holds, but equality does not hold in general, as one can see from the following counterexample.

Example 3. Consider the itemset $X = abcde$ in Example 1, and assume $\tau = 0.15$ and $\gamma = 0.4$. Let us verify that $abcde \in RI_{\tau,\gamma} \backslash \{X \in FC_\tau \mid sup(X) > \gamma * mns_\tau(X) \geq mxs_\tau(X)\}$. Clearly, the rule $b \to acde$ is in $AR_{\tau,\gamma}$, having support 2/6 and confidence 0.5. Moreover, by extending the right-hand side or moving the item b to the right-hand side we get only the rules $b \to acdef$, $\emptyset \to abcde$ and $\emptyset \to abcdef$ of confidence 1/4, 2/6 and 1/6, respectively. Hence, we can conclude that $b \to acde \in RR_{\tau,\gamma}$. On the other hand, $mxs_\tau(X) = 1/6$ and $mns_\tau(X) = 2/6$, so $\gamma * mns_\tau(X) = 0.8/6$ is strictly smaller than $mxs_\tau(X)$. In this case, Algorithm *GenRR* does not work correctly since it does not list the rule $b \to acde$ as being representative.

An alternative counterexample is given in the proof of Lemma 1 below.

3 Characterizing Representative Rules

The goal of pruning off sets that do not give representative rules, by keeping only $RI_{\tau,\gamma}$, cannot be reached using the bounds given, as we have seen that this set comprises all X in FC_τ with $sup(X) \geq \gamma * mns_\tau(X) > mxs_\tau(X)$ but may also include other frequent closed sets X that do not satisfy the condition $\gamma * mns_\tau(X) > mxs_\tau(X)$. We consider two alternatives.

3.1 Closed Sets Instead of Minimal Generators

For closed X, $mns_\tau(X)$ is almost the same thing as the minimal support among all proper subsets of X, or again among all proper closed subsets of X; all these notions coincide when X is its own minimal generator, otherwise they only differ due to the minimal generators of X. Therefore it makes sense to try and exclude the minimal generators of X from consideration. This way, we get another parameter,

$$bmns_\tau(X) = \min(\{sup(Y) \mid Y \in FC_\tau, Y \subset X\} \cup \{\infty\}).$$

The value of $bmns_\tau$ is never smaller than mns_τ as we shall shortly see. Thus, there will be more sets that meet the condition $\gamma * bmns_\tau(X) > mxs_\tau(X)$.

Proposition 4. *The following properties hold.*

- $bmns_\tau(X) = \min(\{sup(Y) \mid Y \in FG_\tau, \overline{Y} \subset X\} \cup \{\infty\})$,
- $mns_\tau(X) \leq bmns_\tau(X)$,
- *if* $X \in FC_\tau \cap FG_\tau$ *then* $mns_\tau(X) = bmns_\tau(X)$,

Proof. We omit the proof of the first two claims because they are straightforward. So, let X be a frequent closed set that is also a minimal generator. If $X = \emptyset$, then $mns_\tau(X) = bmns_\tau(X) = \infty$. Otherwise, let $Y \in FG_\tau$ be such that $Y \subset X$ and $mns_\tau(X) = sup(Y)$. Clearly, $\overline{Y} \in FC_\tau$ and $\overline{Y} \subseteq \overline{X} = X$. Since $X \in FG_\tau$ and $Y \subset X$, $sup(Y) > sup(X)$ and hence $sup(\overline{Y}) > sup(X)$, and therefore $\overline{Y} \subset X$. We get $sup(\overline{Y}) \geq bmns_\tau(X)$ and $mns_\tau(X) \geq bmns_\tau(X)$. Combining it with the fact that $mns_\tau(X) \leq bmns_\tau(X)$ always holds, we conclude that $mns_\tau(X) = bmns_\tau(X)$. □

Unfortunately, the new parameter can still leave out some sets in $RI_{\tau,\gamma}$.

Lemma 1. $RI_{\tau,\gamma} \not\subseteq \{X \in FC_\tau \mid sup(X) > \gamma * bmns_\tau(X) \geq mxs_\tau(X)\}$.

Proof. Let $\mathscr{U} = \{a,b,c\}$ and $\mathscr{D}$ be the dataset containing the following 13 transactions: $t_1 = \cdots = t_8 = abc, t_9 = ab, t_{10} = t_{11} = t_{12} = a, t_{13} = b$; assume $\tau = 0.07$ and $\gamma = 0.7$. One can check that, although $ab \in RI_{\tau,\gamma}$ (since $a \to b \in RR_{\tau,\gamma}$), both $bmns_\tau(ab) = 10/13$ and $mns_\tau(ab) = 10/13$; but $\gamma * mns_\tau(ab) = \gamma * bmns_\tau(ab) = 7/13 < 8/13 = mxs_\tau(ab)$. □

The next construction shows that by using $bmns_\tau$ instead of mns_τ we can even leave out some sets in $RI_{\tau,\gamma}$ that would not have been left out otherwise.

Lemma 2. $RI_{\tau,\gamma} \cap \{X \in FC_\tau \mid sup(X) > \gamma * mns_\tau(X) \geq mxs_\tau(X)\} \not\subseteq \{X \in FC_\tau \mid sup(X) > \gamma * bmns_\tau(X) \geq mxs_\tau(X)\}$.

Proof. Let $\mathscr{U} = \{a,b,c,d,e\}$ and $\mathscr{D}$ be a dataset containing 35 transactions: $t_1 = t_2 = abcde, t_3 = t_4 = t_5 = abcd, t_6 \cdots = t_{20} = a$ and $t_{21} = \cdots t_{35} = b$. Pick $\tau = 0.05$ and $\gamma = 0.75$. Note that $ab \to cd \in RR_{\tau,\gamma}$, and therefore $abcd \in RI_{\tau,\gamma}$. Now, $mns_\tau(abcd) = 5/35$, $bmns_\tau(abcd) = 20/35$, $sup(abcd) = 5/35$ and $mxs_\tau(abcd) = 2/35$. Although $\gamma * mns_\tau(abcd) = 3.5/35 = 0.1$ belongs to the interval $[2/35, 5/35)$, $\gamma * bmns_\tau(abcd) = 15/35$ does not. □

3.2 Minimal Generators of Bounded Support

In order to give a complete characterization for the set $RI_{\tau,\gamma}$, let us first introduce the following notation: for a set X in FC_τ, $mxgs_{\tau,\gamma}(X)$ is the maximal support of those minimal generators that are included in X and are not more frequent than $sup(X)/\gamma$:

$$mxgs_{\tau,\gamma}(X) = \max(\{sup(Y) \mid Y \in FG_\tau, Y \subset X, \gamma * sup(Y) \leq sup(X)\} \cup \{0\}).$$

Note that $mxgs_{\tau,\gamma}(X)$ is either 0, or it is greater than or equal to $sup(X)$. We prove two propositions that explain how we can use this value in order to compute the set $RI_{\tau,\gamma}$ and how to find, given $X \in RI_{\tau,\gamma}$, a subset $X_0 \subset X$ such that $X_0 \to X \backslash X_0 \in RR_{\tau,\gamma}$.

Proposition 5. *The following equality holds.*

$$RI_{\tau,\gamma} = \{X \in FC_\tau \mid \gamma * mxgs_{\tau,\gamma}(X) > mxs_\tau(X)\}.$$

Proof. Let X be an arbitrary set in $RI_{\tau,\gamma}$, and take X_0 in FG_τ such that $X_0 \subset X$ and $X_0 \to X \backslash X_0 \in RR_{\tau,\gamma}$.

We have, on one hand, $conf(X_0 \to X \backslash X_0) \geq \gamma$, and on the other hand, the rule should not be in the cover set of any other rule with confidence greater than γ, i. e. $conf(X_0 \to Z \backslash X_0) < \gamma$ for all $Z \in FC_\tau$ with $Z \supset X$.

That is, $sup(X) \geq \gamma * sup(X_0) > sup(Z)$ for all $Z \in FC_\tau$ with $Z \supset X$. From the first inequality, we deduce that X_0 meets all the conditions in order to be considered for the computation of $mxgs_{\tau,\gamma}(X)$, and therefore, $mxgs_{\tau,\gamma}(X) \geq sup(X_0)$. From the second, we get $\gamma * sup(X_0) > mxs_\tau(X)$. We conclude that $\gamma * mxgs_{\tau,\gamma}(X) > mxs_\tau(X)$.

Conversely, let $X \in FC_\tau$ be such that $\gamma * mxgs_{\tau,\gamma}(X) > mxs_\tau(X)$. It is clear that $mxgs_{\tau,\gamma}(X)$ cannot be 0 (since $mxs_\tau(X) \geq 0$), so

$$\{Y \in FG_\tau \mid Y \subset X, \gamma * sup(Y) \leq sup(X)\} \neq \emptyset.$$

Take $X_0 \in FG_\tau$ to be a set of maximal support that belongs to that set. Therefore, we have $mxgs_{\tau,\gamma}(X) = sup(X_0)$. Since $sup(X_0 \to X \backslash X_0) = sup(X) \geq \tau$ and $conf(X_0 \to X \backslash X_0) = \frac{sup(X)}{sup(X_0)} \geq \gamma$ we deduce that $X_0 \to X \backslash X_0 \in AR_{\tau,\gamma}$. Note that for any $Z \supset X$, $conf(X_0 \to Z \backslash X_0) = \frac{sup(Z)}{sup(X_0)} \leq \frac{mxs_\tau(X)}{sup(X_0)} = \frac{mxs_\tau(X)}{mxgs_{\tau,\gamma}(X)} < \gamma$. Moreover, for any $X_0' \subset X_0$, $sup(X_0') > sup(X_0)$ (since $X_0 \in FG_\tau$) and $\gamma * sup(X_0') > sup(X)$ (due to the choice we have made for X_0). This is why $conf(X_0' \to X \backslash X_0') = \frac{sup(X)}{sup(X_0')} < \gamma$. We conclude that $X_0 \to X \backslash X_0 \in RR_{\tau,\gamma}$ and $X \in RI_{\tau,\gamma}$. □

The previous proposition characterizes unequivocally $RI_{\tau,\gamma}$. Simple arithmetic suffices to check that Proposition 5 identifies exactly the closed sets from which representative rules follow as per Example 2. However, we also need a practical method for identifying the set of representative rules. To this end, we give necessary and sufficient conditions for a subset of an itemset in $RI_{\tau,\gamma}$ to be the left-hand side of a representative rule (see Proposition 6).

Proposition 6. *Let $X \in RI_{\tau,\gamma}$, $c_1 = mxs_\tau(X)/\gamma$, $c_2 = sup(X)/\gamma$ and $X_0 \subset X$. Then $X_0 \to X \backslash X_0 \in RR_{\tau,\gamma}$ if and only if $c_1 < sup(X_0) \leq c_2 < mns_\tau(X_0)$.*

Proof. Consider $X \in RI_{\tau,\gamma}$ and $X_0 \subset X$. Clearly, $X_0 \to X \backslash X_0 \in RR_{\tau,\gamma}$ if and only if the rule $X_0 \to X \backslash X_0$ is in $AR_{\tau,\gamma}$ and does not belong to the cover set of any other rule in $AR_{\tau,\gamma}$. That is equivalent to: $sup(X) \geq \tau$, $\frac{sup(X)}{sup(X_0)} \geq \gamma$, $\frac{sup(X)}{sup(X_0')} < \gamma$ for all $X_0' \subset X$ and $\frac{sup(Z)}{sup(X_0)} < \gamma$ for all $Z \supset X$ that satisfy $sup(Z) \geq \tau$.

Now, it is easy to see that:

- $sup(X) \geq \tau$ always holds because $X \in FC_\tau$,
- $\frac{sup(X)}{sup(X_0)} \geq \gamma \Leftrightarrow sup(X_0) \leq c_2$,
- $\forall X_0' \subset X_0 : \frac{sup(X)}{sup(X_0')} < \gamma \Leftrightarrow \frac{sup(X)}{mns_\tau(X_0)} < \gamma \Leftrightarrow c_2 < mns_\tau(X_0)$,
- $\forall Z \supset X : \left(Z \in F_\tau \Rightarrow \frac{sup(Z)}{sup(X_0)} < \gamma\right) \Leftrightarrow \frac{mxs_\tau(X)}{sup(X_0)} < \gamma \Leftrightarrow c_1 < sup(X_0)$,

which concludes the proof. □

The correctness of Algorithm 1 trivially follows from Propositions 5 and 6.

Algorithm 1 RR Generator

Input: support threshold τ, confidence threshold γ
$F_\tau = \{X \subseteq \mathscr{U} \mid sup(X) \geq \tau\}$
$FC_\tau = \{X \in F_\tau \mid \overline{X} = X\}$
$FG_\tau = \{X \in F_\tau \mid \forall Y \subset X, sup(Y) > sup(X)\}$
for all $X \in FG_\tau$ **do**
    $mns_\tau(X) = \min(\{sup(Y) \mid Y \in FG_\tau, Y \subset X\} \cup \{\infty\})$
end for
$RI_{\tau,\gamma} = \emptyset$
for all $X \in FC_\tau \backslash \{\emptyset\}$ **do**
    $mxs_\tau(X) = \max(\{sup(Z) \mid Z \in FC_\tau, Z \supset X\} \cup \{0\})$
    $mxgs_{\tau,\gamma}(X) = \max(\{sup(Y) \mid Y \in FG_\tau, Y \subset X, \gamma * sup(Y) \leq sup(X)\} \cup \{0\})$
    if $\gamma * mxgs_{\tau,\gamma}(X) > mxs_\tau(X)$ **then**
        add X to $RI_{\tau,\gamma}$
    end if
end for
for all $X \in RI_{\tau,\gamma}$ **do**
    $c_1 = mxs_\tau(X)/\gamma$
    $c_2 = sup(X)/\gamma$
    Ant = $\{X_0 \in FG_\tau \mid X_0 \subset X, c_1 < sup(X_0) \leq c_2 < mns_\tau(X_0)\}$
    for all $X_0 \in$ Ant **do**
        output $X_0 \rightarrow X \backslash X_0$
    end for
end for

3.3 An Algorithm for Different Confidence Thresholds

The disadvantage of Algorithm 1, compared to the one in [Kryszkiewicz, 2001], is that, for a given X in FC_τ, $mxgs_{\tau,\gamma}(X)$ depends on the confidence threshold, and hence it cannot be reused once γ has changed, whereas both $mxs_\tau(X)$ and $mns_\tau(X)$ can be computed only once for a given value of τ and then used for different confidence values. On the other hand, Algorithm 1 is guaranteed not to lose representative rules, whereas the one in [Kryszkiewicz, 2001] risks giving incomplete output, as in our counterexamples above.

Instead of computing $mxgs_{\tau,\gamma}(X)$ for each and every γ, one can find the individual points of the interval $(0,1]$ where $mxgs_{\tau,\gamma}(X)$ changes its value. Indeed, given X in $FC_\tau \backslash \{\emptyset\}$, let $\{Y_1,\ldots,Y_{n[X]}\}$ be the set $\{Y \in FG_\tau \mid Y \subset X\}$ in descending order of support. It is easy to see that

$$mxgs_{\tau,\gamma}(X) = \begin{cases} sup(Y_1), & \text{if } \gamma \leq \frac{sup(X)}{sup(Y_1)}, \\ sup(Y_{i+1}), & \text{if } \gamma \in \left(\frac{sup(X)}{sup(Y_i)}, \frac{sup(X)}{sup(Y_{i+1})}\right], \ i \in \{1,\ldots,n[X]-1\}, \\ 0, & \text{otherwise.} \end{cases}$$

Let us introduce the following notation: for $i \in \{1,\ldots,n[X]\}$, $y_i[X] = sup(Y_i)$ and $p_i[X] = sup(X)/sup(Y_i)$. Moreover, $p_0[X] = 0$. Now, each time a new value of the confidence threshold γ is given, one can decide whether a frequent closed set X is in $RI_{\tau,\gamma}$ by simply retrieving the interval $(p_i[X], p_{i+1}[X]]$ with $i \in \{0,\ldots,n[X]-1\}$ to which γ belongs (recall that in this case $mxgs_{\tau,\gamma}(X) = y_{i+1}[X]$) and then checking whether the inequality $\gamma * y_{i+1}[X] > mxs_\tau(X)$ holds. Note that if no such i exists (that is, whenever γ has a value strictly greater than $p_{n[X]}[X]$), $mxgs_{\tau,\gamma}(X)$ takes the value 0, which makes $\gamma * mxgs_{\tau,\gamma}(X)$ smaller than or equal to $mxs_\tau(X)$.

These ideas are implemented in Algorithms 2 and 3.

Algorithm 2 RR Generator - preprocessing phase

Input: support threshold τ
$F_\tau = \{X \subseteq \mathscr{U} \mid sup(X) \geq \tau\}$
$FC_\tau = \{X \in F_\tau \mid \overline{X} = X\}$
$FG_\tau = \{X \in F_\tau \mid \forall Y \subset X, sup(Y) > sup(X)\}$
for all $X \in FG_\tau$ **do**
  $mns_\tau(X) = \min(\{sup(Y) \mid Y \in FG_\tau, Y \subset X\} \cup \{\infty\})$
end for
for all $X \in FC_\tau \backslash \{\emptyset\}$ **do**
  $mxs_\tau(X) = \max(\{sup(Z) \mid Z \in FC_\tau, Z \supset X\} \cup \{0\})$
  $n[X] = |\{Y \in FG_\tau \mid Y \subset X\}|$
  let $\{Y_1,\ldots,Y_{n[X]}\}$ be the set $\{Y \in FG_\tau \mid Y \subset X\}$ in descending order of support
  for all $i \in \{1,\ldots,n[X]\}$ **do**
    $y_i[X] = sup(Y_i)$
    $p_i[X] = sup(X)/y_i[X]$
  end for
  $p_0[X] = 0$
end for

4 Characterizing the Basis for Closure-Based Redundancy

The results of the previous sections can be extended to find a list of rules such that any other rule in $AR_{\tau,\gamma}$ is redundant with respect to one rule in our list and the set

Algorithm 3 RR Generator - second phase

Input: support threshold τ, confidence threshold γ
$RI_{\tau,\gamma} = \emptyset$
for all $X \in FC_\tau \backslash \{\emptyset\}$ **do**
  if $\exists i \in \{0,\ldots,n[X]-1\}$ such that $\gamma \in (p_i[X], p_{i+1}[X]]$ **then**
    if $\gamma * y_{i+1}[X] > mxs_\tau(X)$ **then**
      add X to $RI_{\tau,\gamma}$
    end if
  end if
end for
for all $X \in RI_{\tau,\gamma}$ **do**
  $c_1 = mxs_\tau(X)/\gamma$
  $c_2 = sup(X)/\gamma$
  Ant = $\{X_0 \in FG_\tau \mid X_0 \subset X, c_1 < sup(X_0) \leq c_2 < mns_\tau(X_0)\}$
  for all $X_0 \in$ Ant **do**
    output $X_0 \rightarrow X \backslash X_0$
  end for
end for

of full-confidence implications. This is exactly the idea behind a basis for closure-based redundancy [Balcázar, 2010a].

Let $\mathscr{B}$ be a set of implications, i. e. rules that hold with confidence 1. Partial rule $X' \rightarrow Y'$ is closure-based redundant relative to $\mathscr{B}$ with respect to $X \rightarrow Y$ if any dataset $\mathscr{D}$ in which all the rules in $\mathscr{B}$ hold with confidence 1 gives $conf(X' \rightarrow Y') \geq conf(X \rightarrow Y)$.

Closure-based redundancy and standard redundancy coincide when the set of implications $\mathscr{B}$ is empty. Knowing the set $\mathscr{B}$ is equivalent to knowing how the closure operator works on each set. If the set of implications is empty, then any subset is closed and all the closure-related argumentations trivialize; in particular, in this case the set of representative rules forms a minimum-size basis.

In any case, we have the following characterization for closure-based redundancy:

Theorem 1 ([Balcázar, 2010a]). *Let $\mathscr{B}$ be a set of exact rules, with associated closure operator mapping each itemset Z to its closure $\overline{Z}$. Let $X' \rightarrow Y'$ be a rule not implied by $\mathscr{B}$, that is, $Y' \not\subset \overline{X'}$, then the following are equivalent:*

1. *$X \subseteq \overline{X'}$ and $X'Y' \subseteq \overline{XY}$,*
2. *The rule $X' \rightarrow Y'$ is closure-based redundant relative to $\mathscr{B}$ with respect to $X \rightarrow Y$.*

Note that $Y' \not\subset \overline{X'}$ is equivalent to saying that $X' \rightarrow Y'$ is not a full implication. One can then analogously define the closure-based cover set of a rule $X \rightarrow Y$ by $\overline{C}(X \rightarrow Y) = \{X' \rightarrow Y' \mid X \subseteq \overline{X'} \text{ and } X'Y' \subseteq \overline{XY}\}$. Accordingly, we must refine the notion of "different" rule since only the closures are relevant: A rule $X' \rightarrow Y'$ is closure-equivalent (again relative to $\mathscr{B}$) to $X \rightarrow Y$ when $\overline{X'} = \overline{X}$ and $\overline{X'Y'} = \overline{XY}$.

The minimum-size basis $\mathscr{B}^*_{\tau,\gamma}$ for closure-based redundancy contains all rules in $AR_{\tau,\gamma}$ of confidence strictly smaller than 1 that are not closure-based redundant with respect to any rule in $AR_{\tau,\gamma}$, unless they are closure-equivalent (see

[Balcázar, 2010a] for details). Again the main property of this basis is that every rule in $AR_{\tau,\gamma}$ is closure-based redundant with a rule in the basis.

Proposition 7. *If a rule is not in the basis, then it is closure-based redundant with respect to a rule in the basis that is not closure-equivalent to it.*

Proof. Indeed, if $X \to Y \backslash X$ is not in the basis, some rule $X' \to Y' \backslash X'$ exists above the confidence and support thresholds for which $X' \subseteq \overline{X}$ and $Y \subseteq \overline{Y'}$, and either $\overline{X'} \neq \overline{X}$ or $\overline{Y'} \neq \overline{Y}$; in turn, this rule is closure-based redundant with a rule in the basis, possibly itself, say $X'' \to Y'' \backslash X''$, so that $X'' \subseteq \overline{X'} \subseteq \overline{\overline{X}} = \overline{X}$ and $Y \subseteq \overline{Y'} \subseteq \overline{\overline{Y''}} = \overline{Y''}$; further, then, $\overline{X''} = \overline{X}$ implies $\overline{X'} = \overline{X}$, and $\overline{Y''} = \overline{Y}$ implies $\overline{Y'} \neq \overline{Y}$. Therefore, if $X \to Y \backslash X$ is not in the basis, then it is closure-based redundant with $X'' \to Y'' \backslash X''$, which is in the basis and is not closure-equivalent to it. □

It is easy to check that, in all rules in this basis, the left-hand sides are also closed sets. We are interested in computing this basis fast. To do that, let $\overline{RI_{\tau,\gamma}}$ be the set of all frequent closed itemsets from which at least one rule for this basis can be obtained.

Proposition 8. *The following equality holds.*

$$\overline{RI_{\tau,\gamma}} = \{X \in FC_\tau \mid \gamma * mxgs_{\tau,\gamma}(X) > mxs_{\tau,\gamma}(X) \text{ and } mxgs_{\tau,\gamma}(X) > sup(X)\}.$$

Proof. Let X be an arbitrary set in $\overline{RI_{\tau,\gamma}}$: there is a basis rule $X_0 \to X \backslash X_0$ for these confidence and support thresholds, where X_0 is a proper closed subset $X_0 \subset X$. Pick a minimal generator X_1 of X_0; as X_0 is closed, $sup(X_1) = sup(X_0) > sup(X)$; as $conf(X_0 \to X \backslash X_0) \geq \gamma$, $\gamma * sup(X_1) = \gamma * sup(X_0) \leq sup(X)$, hence X_1 participates in the computation of $mxgs_{\tau,\gamma}(X)$, so that $mxgs_{\tau,\gamma}(X) \geq sup(X_1) > sup(X)$.

Besides, if there was a proper closed superset Z of X such that $sup(Z) \geq \tau$ and $c(X_0 \to Z \backslash X_0) \geq \gamma$, then the rule $X_0 \to X \backslash X_0$ would not be in the basis due to redundancy with $X_0 \to Z \backslash X_0$. Therefore, the support of any frequent itemset Z with $X \subset Z$ is less than $\gamma * sup(X_0)$. That is, $mxs_{\tau,\gamma}(X) < \gamma * sup(X_0)$. Hence, $\gamma * mxgs_{\tau,\gamma}(X) \geq \gamma * sup(X_1) = \gamma * sup(X_0) > mxs_{\tau,\gamma}(X)$.

Conversely, assume that

$$\gamma * mxgs_{\tau,\gamma}(X) > mxs_{\tau,\gamma}(X) \text{ and } mxgs_{\tau,\gamma}(X) > sup(X)$$

holds for $X \in FC_\tau$. Indeed, $sup(X) < mxgs_{\tau,\gamma}(X)$ implies that this last value is not zero, and that there is at least one itemset $X_1 \in FG_\tau$ such that $X_1 \subset X$ and $\gamma * sup(X_1) \leq sup(X)$. Among these X_1, we pick one with maximum support: $mxgs_{\tau,\gamma}(X) = sup(X_1)$. Let $X_0 = \overline{X_1}$, so $sup(X_0) = sup(X_1) > sup(X)$ and $X_0 \subset X$. Then $conf(X_0 \to X \backslash X_0) = sup(X)/sup(X_0) \geq \gamma * sup(X_1)/sup(X_0) = \gamma$, which implies $X_0 \to X \backslash X_0 \in AR_{\tau,\gamma}$.

Suppose, for a contradiction, that $X_0 \to X \backslash X_0$ is not in the basis. By Proposition 7, it must be closure-based redundant with respect to a rule $Y \to Z \backslash Y$ that is in the basis and is not closure-equivalent to it. Being in the basis implies that $Y,\ Z \in FC_\tau$ (and keep in mind that both X_0 and X are closed as well). By Theorem 1, we have

that $Y \subseteq X_0$ and $X \subseteq Z$, where one of the two inclusions must be proper to ensure closure-inequivalence. If $X \subset Z$, we have that

$$conf(Y \to Z\backslash Y) = \frac{sup(Z)}{sup(Y)} \leq \frac{sup(Z)}{sup(X_0)} \leq \frac{mxs_\tau(X)}{mxgs_{\tau,\gamma}(X)} < \gamma,$$

which is a contradiction with $conf(Y \to Z\backslash Y) \geq \gamma$ as $Y \to Z\backslash Y \in \mathscr{B}^*_{\tau,\gamma} \subseteq AR_{\tau,\gamma}$. The other possibility is that $Z = X$ and $Y \subset X_0$, but $sup(Y) > sup(X_0)$, because $Y \in FC_\tau$, contradicting the maximality of $sup(X_0)$. This finishes the proof. □

Proposition 9. *Let $X \in \overline{RI_{\tau,\gamma}}$, $c_1 = mxs_\tau(X)/\gamma$, and $c_2 = sup(X)/\gamma$. Consider a proper closed subset $X_0 \subset X$. Then $X_0 \to X\backslash X_0 \in \mathscr{B}^*_\gamma$ if and only if $c_1 < \sup(X_0) \leq c_2 < mns_\tau(X_0)$.*

Proof. Consider $X \in \overline{RI_{\tau,\gamma}}$ and a proper closed subset $X_0 \subset X$. The rule $X_0 \to X\backslash X_0$ is in $\mathscr{B}^*_\gamma$ if and only if it meets the support and confidence threshold requirements with respect to τ and γ, it is not a full implication, and is not closure-based redundant with respect to another rule $Y \to Z\backslash Y$.

First of all $sup(X) \geq \tau$, because $X \in \overline{RI_{\tau,\gamma}}$ so it remains to see that:

1. $conf(X_0 \to X\backslash X_0) \geq \gamma$,
2. $conf(Y \to Z\backslash Y) < \gamma$ for any $Y, Z \in FC_\tau$ such that $Y \subseteq X_0$ and $X \subseteq Z$, with at least one of the two inclusions proper.

The first item is equivalent to $sup(X_0) \leq c_2$; for the second item we will divide the proof in two different steps: first, we are going to consider the case where $Y \subset X_0$ and $X \subseteq Z$.

$$\forall Y \subset X_0,\ conf(Y \to Z\backslash Y) < \gamma \iff \frac{sup(X)}{sup(Y)} < \gamma \iff c_2 < mns_\tau(X_0).$$

In a similar way, we obtain that for all Z such that $X \subset Z$ and $Y = X_0$, $conf(Y \to Z\backslash Y) < \gamma$ is equivalent to $c_1 < sup(X_0)$. This finishes the proof. □

All the three algorithms defined so far can be modified to output the set $\mathscr{B}^*_{\tau,\gamma}$ of closure-based irredundant partial rules. These modifications are easy from the results we have proven in this Section, so they are omitted.

5 Empirical Comparison

We have seen that one can find toy examples of datasets in which the output of the algorithm in [Kryszkiewicz, 2001] is incomplete.

We have tested our algorithms on two real-world datasets: the training set part of the UCI Adult US census dataset (see [Asuncion and Newman, 2007]) and a Retail dataset (see [Brijs et al., 1999]).

We have implemented three different algorithms: one for the incomplete heuristic given in [Kryszkiewicz, 2001], one that generates the complete set of representative

rules as described by Algorithm 1, and the last algorithm outputs a complete basis under the notion of closure-based redundancy. In order to get comparable results, all algorithms allow rules with empty antecedent and use the same definition of frequent sets and association rules as given in our preliminaries. We emphasize that, in general, the incomplete heuristic fails to produce a complete basis of representative rules. The code is available at [Balcázar, 2010b].

The first dataset under study, which we refer by the name of Retail, is a market basket data which consists of 88163 transactions over 16470 attributes. In order to preserve the anonymity of the clients, the data has been processed so that each item is represented by a number and each line break separates different customers. For the interested reader, the paper [Brijs et al., 1999] contains more information about this dataset.

Table 2 shows the number of representative rules obtained for different support and confidence thresholds (the seventh column), the cardinality of the output set when mns_τ is used (the fifth column) and the time elapsed in order to obtain them (the sixth and forth columns, respectively). We can see that although for higher support thresholds the output of the algorithms is, most of the times, identical (recall that the output of the algorithm in [Kryszkiewicz, 2001] is always a subset of the whole set of representative rules), lowering both thresholds shows bigger differences.

Table 2 Comparison between *GenRR* and Algorithm 1 on the Retail dataset

Data			*GenRR*		Algorithm 1	
$\lvert FC_\tau \rvert$	Support	Confidence	Time	Rules	Time	Rules
7573	0.1%	0.9	0.015	248	0.013	248
		0.8	0.013	643	0.013	652
		0.7	0.028	1978	0.026	1990
19115	0.05%	0.9	0.036	670	0.022	670
		0.8	0.073	2228	0.041	2229
		0.7	0.123	6029	0.083	6039

Dataset Adult is a transactional version of the training set part of the UCI census dataset Adult US (see [Asuncion and Newman, 2007]); it consists of 32561 transactions over 269 items. On the Adult dataset, we see the same trend in the behavior of both algorithms. Note that in this case there are significant differences between the output of the algorithm in [Kryszkiewicz, 2001] and the set of all representative rules (Table 3). For example, for support and confidence thresholds of 0.05 and 0.7, respectively, more than half of the rules are lost.

As an example, in the case the thresholds for support and confidence are 1% and 0.70, respectively, there are a total of 6867 representative rules, among which 3408 are lost when using *mns* or *bmns* (four of them listed in bold, the rest of the rules are given as an example):

[c:0.75, s:1.03] Private White age: 41 ⇒ Male,
[c:0.82, s:2.21] Never-married Unmarried ⇒ <=50K USA,
[c:0.70, s:1.47] <=50K Assoc-acdm White ⇒ Private,

Table 3 Comparison between *GenRR* and Algorithm 1 on the Adult dataset

Data			*GenRR*		Algorithm 1	
$\|FC_\tau\|$	Support	Confidence	Time	Rules	Time	Rules
11920	1%	0.9	0.147	6578	0.176	7436
		0.8	0.130	4827	0.148	7379
		0.7	0.096	3459	0.141	6867
27444	0.5%	0.9	0.391	15208	0.380	17573
		0.8	0.298	11516	0.417	18190
		0.7	0.263	8241	0.382	16779

[c:0.75, s:3.74] Own-child Private hours-per-week: 40 $\Rightarrow$ <=50K Never-married USA,
[c:0.75, s:3.74] Never-married Own-child USA hours-per-week: 40 $\Rightarrow$ <=50K Private,
[c:0.87, s:1.03] Male Private age: 41 $\Rightarrow$ White
[c:0.75, s:1.03] Private White age: 41 $\Rightarrow$ Male
[c:0.86, s:7.07] Exec-managerial Private $\Rightarrow$ USA White
[c:0.73, s:1.04] Craft-repair Divorced $\Rightarrow$ Male USA White
[c:0.75, s:1.68] Not-in-family hours-per-week: 50 $\Rightarrow$ <=50K

As mentioned in the beginning of this section, we have run experiments in order to see the performance of our algorithm that finds a basis under closed-based redundancy conditions. The results are in Tables 4 and 5. Notice that in this case the times are significantly lower.

Table 4 Algorithm for Basis $\mathscr{B}^*_{\tau,\gamma}$ (Retail dataset)

Support	Confidence	Time	Rules
0.1%	0.9	0.006	233
	0.8	0.007	643
	0.7	0.013	1984
0.05%	0.9	0.029	549
	0.8	0.024	2139
	0.7	0.044	6039

Table 5 Algorithm for Basis $\mathscr{B}^*_{\tau,\gamma}$ (Adult dataset)

Support	Confidence	Time	Rules
1%	0.9	0.093	7103
	0.8	0.086	7205
	0.7	0.082	6662
0.5%	0.9	0.243	16457
	0.8	0.250	17531
	0.7	0.233	16085

We have run the experiments on an Intel Core i3-330M @ 2,13GHz machine with 4 GB of RAM running under Microsoft Windows 7 Professional (64 bits). The running time of all algorithms were between 6 and 123 milliseconds in the case

of the Retail dataset and between 82 and 417 milliseconds for the Adult dataset. Algorithm 1 correctly outputs all representative rules at the cost of being sometimes slower than the possibly incomplete algorithm of Kryszkiewicz but, in our tests, the difference was rather irrelevant since the time needed to print the results on screen (a device slower than the CPU) still dominates the process.

It must be noted that the quantity of representative rules may decrease at lower confidence or support thresholds. This phenomenon has been observed and explained before (see [Balcázar, 2010a]) and is caused by powerful rules of a given confidence, say 0.8, that are filtered out at higher thresholds, leaving therefore many other rules as representative, but that force all of these out of the representative rules set as they become redundant when the confidence threshold gets below 0.8 and lets the powerful rule in.

6 Conclusions

We have proposed an alternative (complete) solution for the generation of the set of all representative rules defined in [Kryszkiewicz, 1998b] (see Algorithm 1); we have also shown that the original algorithm was incomplete. Our approach, which seems to requiere more operations than the one in [Kryszkiewicz, 2001], has the advantage of being guaranteed to output the whole set of representative rules.

On the other hand, one of its main drawbacks is that we cannot reuse the precomputed values of the parameters once the user changes the confidence threshold. Our proposal for fixing this problem involves dividing the process into two phases (see Algorithm 2 and Algorithm 3). As a conclusion, depending on whether one is interested in getting complete results or getting them faster, it is more convenient to use Algorithm 1 or the algorithm in [Kryszkiewicz, 2001].

We have also extended our approach to the similar but different basis corresponding to closure-based redundancy. Tests were performed in other to confirm that the algorithm is significantly faster than the previous two.

References

Aggarwal and Yu, 2001. Aggarwal, C. C. and Yu, P. S. (2001). A new approach to online generation of association rules. *IEEE Transactions on Knowledge and Data Engineering*, 13(4):527–540.

Agrawal et al., 1996. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press.

Asuncion and Newman, 2007. Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Balcázar, 2010a. Balcázar, J. L. (2010a). Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science*, 6(2:3):1–33.

Balcázar, 2010b. Balcázar, J. L. (2010b). Slatt023. `https://code.google.com/p/slatt/downloads/list.`

Brijs et al., 1999. Brijs, T., Swinnen, G., Vanhoof, K., and Wets, G. (1999). Using association rules for product assortment decisions: A case study. In *Knowledge Discovery and Data Mining*, pages 254–260.

Ceglar and Roddick, 2006. Ceglar, A. and Roddick, J. F. (2006). Association mining. *ACM Comput. Surv.*, 38(2).

Cristofor and Simovici, 2002. Cristofor, L. and Simovici, D. A. (2002). Generating an informative cover for association rules. In *Proc. of the 2002 IEEE International Conference on Data Mining (ICDM)*, pages 597–600. IEEE Computer Society.

Hamrouni et al., 2008. Hamrouni, T., Ben Yahia, S., and Mephu Nguifo, E. (2008). Succinct minimal generators: Theoretical foundations and applications. *Int. J. Found. Comput. Sci.*, 19(2):271–296.

Kryszkiewicz, 1998a. Kryszkiewicz, M. (1998a). Fast discovery of representative association rules. In Polkowski, L. and Skowron, A., editors, *Proc. of the 1st International Conference on Rough Sets and Current Trends in Computing (RSCTC)*, volume 1424 of *Lecture Notes in Artificial Intelligence*, pages 214–221. Springer-Verlag.

Kryszkiewicz, 1998b. Kryszkiewicz, M. (1998b). Representative association rules. In Wu, X., Ramamohanarao, K., and Korb, K. B., editors, *Proc. of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 1394 of *Lecture Notes in Artificial Intelligence*, pages 198–209. Springer-Verlag.

Kryszkiewicz, 2001. Kryszkiewicz, M. (2001). Closed set based discovery of representative association rules. In Hoffmann, F., Hand, D. J., Adams, N. M., Fisher, D. H., and Guimarães, G., editors, *Proc. of the 4th International Symposium on Intelligent Data Analysis (IDA)*, volume 2189 of *Lecture Notes in Computer Science*, pages 350–359. Springer-Verlag.

Kryszkiewicz, 2002. Kryszkiewicz, M. (2002). Concise representations of association rules. In Hand, D. J., Adams, N. M., and Bolton, R. J., editors, *Proc. of the ESF Exploratory Workshop on Pattern Detection and Discovery*, volume 2447 of *Lecture Notes in Computer Science*, pages 92–109. Springer-Verlag.

Luxenburger, 1991. Luxenburger, M. (1991). Implications partielles dans un contexte. *Mathématiques et Sciences Humaines*, 29:35–55.

Pasquier et al., 2005. Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., and Lakhal, L. (2005). Generating a condensed representation for association rules. *J. Intell. Inf. Syst.*, 24(1):29–60.

Phan-Luong, 2001. Phan-Luong, V. (2001). The representative basis for association rules. In Cercone, N., Lin, T. Y., and Wu, X., editors, *ICDM*, pages 639–640. IEEE Computer Society.

Zaki, 2004. Zaki, M. J. (2004). Mining non-redundant association rules. *Data Min. Knowl. Discov.*, 9(3):223–248.