New Exact Concise Representation of Rare Correlated Patterns: Application to Intrusion Detection

Souad Bouasker[1], Tarek Hamrouni[1], and Sadok Ben Yahia[1,2]

[1] LIPAH, Computer Science Department, Faculty of Sciences of Tunis, Tunis, Tunisia
[2] Institut TELECOM, TELECOM SudParis, UMR 5157 CNRS SAMOVAR, France

Abstract. During the last years, many works focused on the exploitation of rare patterns. In fact, these patterns allow conveying knowledge on unexpected events. Nevertheless, a main problem is related to their very high number and to the low quality of several mined rare patterns. In order to overcome these limits, we propose to integrate the correlation measure *bond* aiming at only mining the set of rare *correlated* patterns. A characterization of the resulting set is then detailed, based on the study of constraints of different natures induced by the rarity and the correlation. In addition, based on the equivalence classes associated to a closure operator dedicated to the *bond* measure, we propose a new exact concise representation of rare correlated patterns. We then design the new RCPRMINER algorithm allowing an efficient extraction of the proposed representation. The carried out experimental studies prove the compactness rate offered by our approach. We also design an association rules based classifier and we prove its effectiveness in the context of intrusion detection.

Keywords: Concise representation, Constraint, Rarity, Correlation, Closure operator, Equivalence class.

1 Introduction and Motivations

Recently, rare pattern mining has been proved to be of actual added value in many application fields such as the intrusion detection, the analysis of criminal data, the pharmacovigilance, etc. [7]. In fact, rare patterns can identify unexpected events or exceptions [15], since they have a very low frequency in the data base. In practice, the exploitation of rare patterns is hampered by the high number and the low quality of the extracted rare patterns. Thus, an extracted rare pattern may not represent any useful information whenever it is composed only by items among which there is no *semantic* link. In this situation, integrating correlation measures would be of benefit by only mining *Rare correlated patterns*. These latter patterns offer a strong semantic link among their items. Indeed, an interesting rare pattern is that which appears a very small number of times in the database but has items that are strongly linked w.r.t. a correlation metric.

In this paper, we focus on the extraction of an exact concise representation of rare correlated patterns w.r.t. to the *bond* correlation measure [10]. This measure is redefined in this work as the ratio between the conjunctive support of a pattern and its disjunctive support. Indeed, although used in many works under various names like *extended Jaccard measure*, *coherence*, *Tanimoto coefficient*, the link between the expression of this

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 61–72, 2012.

measure and the disjunctive support in the denominator part has never been established in the literature. Our choice of this measure is motivated by the theoretical framework presented in [10,14], in addition to the structural study that was done in [12]. Furthermore, it has been proved in [14] that the *bond* measure fulfills the theoretical properties that any measure of quality dedicated to rare association rule should have. Moreover, the authors in [12] proposed a generic approach for correlated patterns mining based on the *bond* measure. Note, however, that the study of rare correlated patterns was not previously carried out in the literature.

From the computational point of view, the integration of the *bond* measure within the mining process of rare patterns is a very challenging task. Indeed, the correlated patterns associated to the *bond* measure verify an anti-monotone constraint and then induce an *order ideal* [5] in the pattern lattice. In opposition to this, rare patterns form an *order filter* [5] in the pattern lattice since they fulfill a monotone constraint. Therefore, the set of rare correlated patterns result from the intersection of two theories [9] induced by the constraints of correlation and rarity. The set of rare correlated patterns is then more complicated to be mined than any set of patterns induced by one or more constraints of the same nature [4]. We thus provide in this paper a thorough characterization of this set of patterns based on the notion of equivalence class. In our case, equivalence classes are induced by the closure operator associated to the *bond* measure.

To the best of our knowledge, there is no previous study in the literature that has been dedicated to the extraction of a concise representation of patterns fulfilling both the rarity and the correlation constraints. Worth of mention that the new proposed approach is generic and can then be applied to any set of rare correlated pattern according to any correlation measure which shares the same structural properties as the *bond* measure, *e.g.*, the *all-confidence* measure [10].[1]

The remainder of the paper is organized as follows: Section 2 presents basic notions used throughout this work. In Section 3, we characterize the set of all rare correlated patterns by studying the associated constraints. We also introduce the associated new exact concise representation. Section 4 is dedicated to the description of the RCPRMINER algorithm allowing the extraction of the proposed representation. The empirical studies are provided in Section 5. Section 6 illustrates the application of our approach in intrusion detection. The conclusion and perspectives are sketched in Section 7.

2 Basic Notions

We start by presenting the key notions related to our work. We first define a dataset.

Definition 1. (***Dataset***) *A dataset is a triplet* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ *where* $\mathcal{T}$ *and* $\mathcal{I}$ *are, respectively, a finite set of transactions and items, and* $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$ *is a binary relation between the transaction set and the item set. A couple* $(t, i) \in \mathcal{R}$ *denotes that the transaction* $t \in \mathcal{T}$ *contains the item* $i \in \mathcal{I}$.

In this work, we are mainly interested in itemsets as a class of patterns. The two main kinds of support a pattern can have are defined as follows, for any non-empty pattern I:
- **Conjunctive support**: $Supp(\wedge I) = |\{t \in \mathcal{T} \mid (\forall\, i \in I, (t, i) \in \mathcal{R})\}|$
- **Disjunctive support**: $Supp(\vee I) = |\{t \in \mathcal{T} \mid (\exists\, i \in I, (t, i) \in \mathcal{R})\}|$

[1] Mathematically equivalent to the *h-confidence* measure [16].

Table 1. An example of a dataset

	A	B	C	D	E
1	×		×	×	
2		×	×		×
3	×	×	×		×
4		×			×
5	×	×	×		×

Example 1. Let us consider the dataset given by Table 1. We have *Supp*($\wedge$AD) = $|\{1\}|$ = 1 and *Supp*($\vee$AD) = $|\{1, 3, 5\}|$ = 3.[2]

We distinguish, given a minimum support threshold [1], between *frequent* and *rare* patterns. These latter patterns are defined as follows.

Definition 2. (***Rare patterns***) *The set of rare patterns is defined by:* $\mathcal{RP} = \{I \subseteq \mathcal{I} | Supp(\wedge I) < minsupp\}$.

Among the elements of $\mathcal{RP}$, we distinguish the smallest rare patterns according to set-inclusion relation. These patterns constitute the set $\mathcal{M}in\mathcal{RP}$ defined as follows:

Definition 3. (***Minimal rare patterns***) *The* $\mathcal{M}in\mathcal{RP}$ *set of minimal rare pattern is composed by rare patterns having no rare proper subset. It is equal to:* $\mathcal{M}in\mathcal{RP} = \{I \in \mathcal{RP} | \forall\ I_1 \subset I: I_1 \notin \mathcal{RP}\}$.

Example 2. Let us consider the dataset given by Table 1 for *minsupp* = 4. We have, for example, the pattern BC $\in \mathcal{RP}$ since *Supp*($\wedge$BC) = 3 $<$ 4. We also have the pattern BC $\in \mathcal{M}in\mathcal{RP}$ since *Supp*($\wedge$BC) = 3 $<$ 4 and, on the other hand, *Supp*($\wedge$B) = *Supp*($\wedge$C) = 4. In this case, $\mathcal{M}in\mathcal{RP}$ = {A, D, BC, CE}.

In the following, we define *monotone* and *anti-monotone* constraints [4,11].

Definition 4. (***Monotone/Anti-Monotone constraint***) *Let* Q *be a constraint,*
- Q *is anti-monotone if* $\forall\ I \subseteq \mathcal{I}$, $\forall\ I_1 \subseteq I$: I *fulfills* $Q \Rightarrow I_1$ *fulfills* Q
- Q *is monotone if* $\forall\ I \subseteq \mathcal{I}$, $\forall\ I_1 \supseteq I$: I *fulfills* $Q \Rightarrow I_1$ *fulfills* Q

The constraint of rarity is a monotone constraint, *i.e.*, $\forall\ I,\ I_1 \subseteq \mathcal{I}$, if $I_1 \supseteq I$ and *Supp*($\wedge I$) $<$ *minsupp*, then *Supp*($\wedge I_1$) $<$ *minsupp* since *Supp*($\wedge I_1$) $\leq$ *Supp*($\wedge I$). Thus, it induces an *order filter* [5] on the set of all the subsets of $\mathcal{I}$, $\mathcal{P}(\mathcal{I})$. Worth of mention that the frequency constraint induces an *order ideal* [5].

Definition 6 presents the set of correlated patterns according to the *bond* measure [10] which is redefined here as given in Definition 5.

Definition 5. (***The bond measure***) *The bond measure of a non-empty pattern* $I \subseteq \mathcal{I}$ *is defined as follows:*

$$bond(I) = \frac{Supp(\wedge I)}{Supp(\vee I)}$$

Definition 6. (***Correlated patterns***) *Considering a minimum correlation threshold minbond, the set* $\mathcal{CP}$ *of correlated patterns is equal to:* $\mathcal{CP} = \{I \subseteq \mathcal{I} \mid bond(I) \geq minbond\}$.

[2] We use a separator-free form for the sets, *e.g.*, AD stands for the set of items {A, D}.

The constraint of correlation is an anti-monotone constraint, *i.e.*, $\forall\ I,\ I_1 \subseteq \mathcal{I}$, if $I_1 \subseteq I$, then $bond(I_1) \geq bond(I)$. Therefore, the set $\mathcal{CP}$ of correlated patterns forms an *order ideal* [5] on $\mathcal{P}(\mathcal{I})$.

In the following, we will need the set composed by the maximal correlated patterns which constitute the positive border of correlated patterns. This set is defined as follows:

Definition 7. (***Maximal correlated patterns***) *The set of maximal correlated patterns, denoted $\mathcal{M}ax\mathcal{CP}$, is composed by correlated patterns having no correlated proper superset, i.e., $\mathcal{M}ax\mathcal{CP} = \{I \in \mathcal{CP} | \forall\ I_1 \supset I: I_1 \notin \mathcal{CP}\}$.*

Example 3. Consider the dataset illustrated by Table 1. For *minbond* = 0.2, we have $bond(\texttt{BCE}) = \frac{3}{5} = 0.6 \geq 0.2$. Therefore, the pattern BCE is a correlated one. In addition, whatever the strict superset of BCE, this superset is not correlated. In this case, we have $\mathcal{M}ax\mathcal{CP} = \{\texttt{ACD}, \texttt{ABCE}\}$.

Now we focus on the closure operator associated to the *bond* measure.

Definition 8. (***The operator*** f_{bond}) *The closure operator f_{bond}: $\mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{I})$ associated to the bond measure is defined as follows: $f_{bond}(I) = I \cup \{i \in \mathcal{I} \setminus I \mid bond(I) = bond(I \cup \{i\})\}$.*

The closure of a pattern I by f_{bond}, *i.e.* $f_{bond}(I)$, corresponds to the maximal set of items containing I and sharing the same *bond* value with I. It can be easily proven that $f_{bond}(I)$ is equal to the intersection of its conjunctive closure f_c and its disjunctive one f_d. Indeed, according to the framework proposed in [13], *bond* measure is a condensable function based on both preserving functions, namely the conjunctive support and the disjunctive support.

Example 4. Consider the dataset illustrated by Table 1. For *minbond* = 0.2, we have $f_{bond}(\texttt{AB}) = \texttt{ABCE}$ since C and E preserve the *bond* value of AB.

We focus now on the equivalence classes induced by the f_{bond} closure operator.

Definition 9. (***Equivalence class associated to the closure operator*** f_{bond}) *An equivalence class associated to the closure operator f_{bond} is composed by all the patterns having the same closure by the operator f_{bond}. Let $[I]$ be the equivalence class to which belongs the pattern I. $[I]$ is formally defined as follows: $[I] = \{I_1 |\ f_{bond}(I) = f_{bond}(I_1)\}$.*

In each class, all the patterns share the same *bond* value as well as the same conjunctive, disjunctive, and negative supports. Therefore, all the patterns belonging to the same class, induced by f_{bond}, appear exactly in the same transactions. Besides, these patterns characterize the same set of transactions. Indeed, each transaction necessarily contains a non-empty subset of each pattern of the class.

In the next section, we will carry out a detailed study of the rare correlated patterns.

3 Characterization of the Rare Correlated Patterns

3.1 Definition and Properties

The set of rare correlated patterns is defined as follows:

Definition 10. (***Rare correlated patterns***) *Considering the support threshold minsupp and the correlation threshold minbond, the set of rare correlated patterns, denoted $\mathcal{RCP}$, is equal to:* $\mathcal{RCP} = \{I \subseteq \mathcal{I} \mid Supp(\wedge I) < minsupp \text{ and } bond(I) \geq minbond\}$.

Example 5. Consider the dataset illustrated by Table 1. For *minsupp* = 4 and *minbond* = 0.2, the set $\mathcal{RCP}$ consists of the following patterns where each triplet represents the pattern, its conjunctive support value and its *bond* value: $\mathcal{RCP}$ = {(A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AB, 2, $\frac{2}{5}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (AE, 2, $\frac{2}{5}$), (BC, 3, $\frac{3}{5}$), (CD, 1, $\frac{1}{4}$), (CE, 3, $\frac{3}{5}$), (ABC, 2, $\frac{2}{5}$), (ABE, 2, $\frac{2}{5}$), (ACD, 1, $\frac{1}{4}$), (ACE, 2, $\frac{2}{5}$), (BCE, 3, $\frac{3}{5}$), (ABCE, 2, $\frac{2}{5}$)}. This set is depicted by Figure 1. The support shown at the top left of each frame represents the conjunctive support. As shown in Figure 1, the rare correlated patterns are then localized below the border shown in red of the anti-monotone constraint of correlation, composed by the elements of $\mathcal{M}axC\mathcal{P}$, and over the border shown in black of the monotone constraint of rarity, composed by the elements of $\mathcal{M}in\mathcal{RP}$.

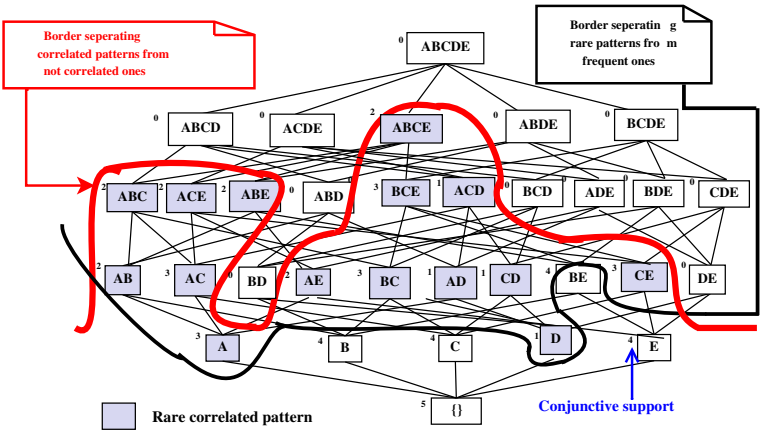

Fig. 1. Localization of the rare correlated patterns for *minsupp* = 4 and *minbond* = 0.2

Therefore, the localization of the rare correlated patterns is much more complex than the localization of theories corresponding to constraints of the same nature.

Interestingly enough, the use of *bond* allows improving the quality of mined patterns by only retaining those containing strongly correlated items. However, the number of these patterns is not necessarily reduced which may hamper its practical use. To losslessly reduce such amount of information, we propose to define a new exact concise representation of the $\mathcal{RCP}$ set. An exact representation of rare correlated patterns should determine, for an arbitrary pattern, whether it is rare correlated or not. If it is a rare correlated one, then this representation must allow faithfully deriving the values of its support and its *bond* measure. In this respect, the proposed representation in this work will be shown to be perfect in the sense that its size never exceeds that of the whole set of rare correlated patterns. It also allows a better exploitation and management of the

extracted knowledge. In addition, since the representation, that we introduce, is lossless, it allows to derive, whenever of need, the whole set of rare correlated patterns.

The new exact concise representation of rare correlated patterns is based on the notion of equivalence class. Equivalence classes allow us to only keep track of non-redundant patterns. Indeed, we retain for each class only the maximal and the minimal ones. The next subsection details our approach.

3.2 Characterization of the Rare Correlated Equivalence Classes

Based on Definition 9, the elements of the same equivalence class have the same behavior w.r.t. both the correlation and the rarity constraints. In fact, for a correlated equivalence class, *i.e.* a class which contains correlated patterns, all of them could be rare or frequent. The application of f_{bond} then provides a more selective process to only extract representative rare correlated patterns of each class. The $\mathcal{RCP}$ set of rare correlated patterns is then split into disjoint equivalence classes – the rare correlated equivalence classes – in which the closed pattern is the largest one w.r.t. the set-inclusion relation. The smallest, w.r.t. the set-inclusion relation, patterns in a class are the minimal rare correlated patterns. The set of these particular patterns are formally defined as follows:

Definition 11. (***Closed rare correlated patterns***) *The $\mathcal{CRCP}$ set of closed rare correlated patterns is equal to:* $\mathcal{CRCP} = \{I \in \mathcal{RCP} | \forall\ I_1 \supset I$: $bond(I) > bond(I_1)\}$.

Definition 12. (***Minimal rare correlated patterns***) *The $\mathcal{MRCP}$ set of minimal rare correlated patterns is equal to:* $\mathcal{MRCP} = \{I \in \mathcal{RCP} | \forall\ I_1 \subset I$: $bond(I) < bond(I_1)\}$.

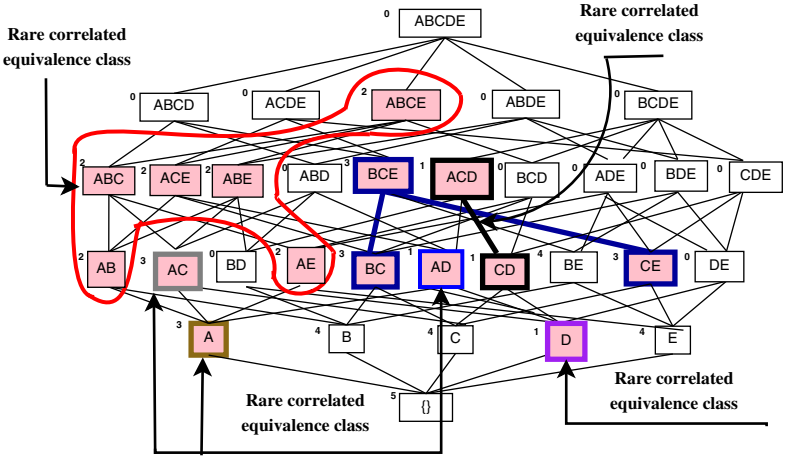

Fig. 2. An example of rare correlated equivalence classes for *minsupp* = 4 and *minbond* = 0.2

Example 6. Consider the dataset $\mathcal{D}$ illustrated by Table 1. For *minsupp* = 4 and *minbond* = 0.2, Figure 2 shows the rare correlated equivalence classes. We have, the set $\mathcal{CRCP}$ = {A, D, AC, AD, ACD, BCE and ABCE}. Whereas, the set $\mathcal{MRCP}$ is equal to: $\mathcal{MRCP}$ = {A, D, AB, AC, AD, AE, BC, CD and CE}. As shown by Figure 2, the patterns A, D, AC and AD are closed and at the same time minimal. Their equivalence classes then contain a unique element.

Before introducing our new concise representation, it is worth mentioning that the notions of closed patterns and minimal generators were also simultaneously used in [8] in order to offer a lossless concise representation of frequent itemsets. Now, based on the two previous sets, we define our new exact concise representation $\mathcal{RCPR}$.[3]

Definition 13. (***The*** $\mathcal{RCPR}$ ***representation***) *The* $\mathcal{RCPR}$ *representation is equal to:* $\mathcal{RCPR} = \mathcal{CRCP} \cup \mathcal{MRCP}$.

Example 7. Consider the dataset illustrated by Table 1 for *minsupp* = 4 and *minbond* = 0.2. According to the previous example, we have the $\mathcal{RCPR}$ representation composed by: (A, 3, $\frac{3}{3}$), (D, 1, $\frac{1}{1}$), (AB, 2, $\frac{2}{5}$), (AC, 3, $\frac{3}{4}$), (AD, 1, $\frac{1}{3}$), (AE, 2, $\frac{2}{5}$), (BC, 3, $\frac{3}{5}$), (CD, 1, $\frac{1}{4}$), (CE, 3, $\frac{3}{5}$), (ACD, 1, $\frac{1}{4}$), (BCE, 3, $\frac{3}{5}$) and (ABCE, 2, $\frac{2}{5}$).

The following theorem proves that the $\mathcal{RCPR}$ representation is a lossless concise representation of the $\mathcal{RCP}$ set. In this respect, both sets $\mathcal{MRCP}$ and $\mathcal{CRCP}$ composing $\mathcal{RCPR}$ are required for the exact regeneration of the set $\mathcal{RCP}$. The elements of the former set are indeed required for ensuring the rarity property of an arbitrary pattern. While the latter set is used for checking the correlation property and for exactly deriving its *bond* and support values.

Theorem 1. *The* $\mathcal{RCPR}$ *representation is an exact concise representation of the* $\mathcal{RCP}$ *set of rare correlated patterns.*

Proof. Let $I \subseteq \mathcal{I}$. We distinguish between three different cases:

a) If $I \in \mathcal{RCPR}$, then I is a rare correlated pattern and we have its support and its *bond* values in the representation.

b) If $\nexists\, J \in \mathcal{RCPR}$ such that $J \subseteq I$ or $\nexists\, Z \in \mathcal{RCPR}$ such that $I \subseteq Z$, then $I \notin \mathcal{RCP}$ since I does not belong to any rare correlated equivalence class.

c) $I \in \mathcal{RCP}$. Indeed, J and Z exist (otherwise I fulfills the conditions of the case b). Thus, I is correlated since it is included in a correlated pattern, namely Z. It is also rare since it contains a rare pattern, namely J. In this case, it is sufficient to localize the f_{bond} closure of I, say F. The closed pattern F belongs to $\mathcal{RCPR}$ since I is rare correlated and $\mathcal{RCPR}$ includes the $\mathcal{CRCP}$ set of closed rare correlated patterns. Therefore, $F = min_{\subseteq}\{I_1 \in \mathcal{RCPR} |\ I \subseteq I_1\}$. Since f_{bond} preserves the *bond* value and the conjunctive support, we then have: *bond*(I) = *bond*(F) and *Supp*($\wedge I$) = *Supp*($\wedge F$). ♢

Example 8. Consider the $\mathcal{RCPR}$ representation illustrated by Example 7. Let us consider each case separately. The pattern AD $\in \mathcal{RCPR}$. Thus, we have its support equal to 1 and its *bond* value equal to $\frac{1}{3}$. Although the pattern BE is included in two patterns from the $\mathcal{RCPR}$ representation, namely BCE and ABCE, BE $\notin \mathcal{RCP}$ since no element of $\mathcal{RCPR}$ is included in BE. Consider now the pattern ABC. There are two patterns of $\mathcal{RCPR}$ which allow determining that the pattern ABC is a rare correlated one, namely AB and ABCE, since AB $\subseteq$ ABC $\subseteq$ ABCE. The smallest pattern in $\mathcal{RCPR}$ which cover ABC, *i.e.* its closure, is ABCE. Then, *bond*(ABC) = *bond*(ABCE) = $\frac{2}{5}$, and *Supp*($\wedge$ABC) = *Supp*($\wedge$ABCE) = 2.

[3] $\mathcal{RCPR}$ stands for **R**are **C**orrelated **P**attern **R**epresentation.

The proof of Theorem 1 clearly highlights that it is straightforward the way queries over of the proposed representation would be carried out w.r.t. a given arbitrary pattern as well as the derivation of the whole set of rare correlated patterns. It is also important to mention that the $\mathcal{RCPR}$ representation is a *perfect cover* of the $\mathcal{RCP}$ set, *i.e.*, the size of $\mathcal{RCPR}$ *never exceeds* that of the $\mathcal{RCP}$ set whatever the dataset and the used *minsupp* and *minbond* values. It is in fact always true that $(\mathcal{CRCP} \cup \mathcal{MRCP}) \subseteq \mathcal{RCP}$.

In the following, we introduce the RCPRMINER algorithm dedicated to the extraction of the $\mathcal{RCPR}$ representation.

4 The RCPRMINER Algorithm

The pseudo-code of RCPRMINER is shown by Algorithm 1. RCPRMINER is a levelwise algorithm which takes as an input a dataset $\mathcal{D}$, a minimum support threshold *minsupp* and a minimum correlation threshold *minbond*. This algorithm allows the determination of the $\mathcal{MRCP}$ and the $\mathcal{CRCP}$ sets which constitute the $\mathcal{RCPR}$ representation.

Algorithm 1. RCPRMINER

Data: A dataset $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$, *minbond*, and *minsupp*.
Results: The exact concise representation $\mathcal{RCPR} = \mathcal{MRCP} \cup \mathcal{CRCP}$.

```
1  Begin
2    RCPR := ∅; Cand_0 := {∅};
3    /* The first step */
4    MaxCP := MAXCP_EXTRACTION(D, minbond);
5    /* The second step */
6    MaxCFP := {X ∈ MaxCP | X.ConjS ≥ minsupp} /* X.ConjS denotes the
     conjunctive support of X */;
7    MaxRCP := {X ∈ MaxCP | X.ConjS < minsupp} ;
8    PCand_1 := {i | i ∈ I} /* PCand_n stands for Potential Candidates of size n */;
9    While (PCand_n ≠ ∅) Do
10     /* Pruning of potential candidate patterns */
11     Cand_n := PCand_n \ {X_n ∈ PCand_n | (∃ Z ∈ MaxCFP: X_n ⊆ Z) or (∄
       Z ∈ MaxRCP: X_n ⊆ Z) or (∃ Y_{n-1} ⊂ X_n: Y_{n-1} ∉ Cand_{n-1})};
12     /* Determination of the minimal rare correlated patterns of size n and
       computation of their closures */
13     RCPR := RCPR ∪ MRCP_CRCP_COMPUTATION(D, Cand_n, minsupp);
14     n := n +1;
15     PCand_n := APRIORI-GEN(Cand_{n-1}) ;
16   Return RCPR;
17 End
```

The RCPRMINER algorithm mainly operates in two steps. The first step consists in extracting the maximal correlated patterns from the extraction context through the invocation of the dedicated MAXCP_EXTRACTION procedure (*cf.* Line 4). The second step consists in integrating the constraint of rarity and the obtained maximal correlated patterns in a mining process of $\mathcal{RCPR}$. In this situation, the set $\mathcal{PC}and_n$ of potential candidates of size n is obtained using the classical APRIORI-GEN procedure (*cf.* Line 15) from the retained candidates of size (n - 1). Once obtained, the set elements

of $\mathcal{PC}and_n$ are pruned (*cf.* Line 11) using several pruning strategies to yield the set $\mathcal{C}and_n$. The used pruning strategies are as follows:

(***i***) **The pruning of the candidates which are included in a maximal correlated frequent pattern.**

(***ii***) **The pruning of the candidates which are not included in a maximal rare correlated pattern.**

(***iii***) **The pruning based on the order ideal of the minimal correlated patterns**.

Recall that the set of minimal correlated patterns induces an order ideal property. Therefore, every minimal correlated candidate, having a non minimal correlated subset, will be pruned since it will not be a minimal correlated pattern. In this respect, within the MRCP_CRCP_COMPUTATION procedure (*cf.* Line 13), whose pseudo-code is omitted here for lack of available space, the minimal rare correlated patterns are determined among the retained candidates in $\mathcal{C}and_n$ (*cf.* Line 11). This is done by comparing their *bond* values to those of their respective immediate subsets. Then, minimal rare correlated patterns are inserted into the $\mathcal{MRCP}$ set. Their closures are after that computed according to the f_{bond} closure operator, and then, inserted in the $\mathcal{CRCP}$ set. From the computational point of view, it is important to mention that the localization of the border composed by the maximal correlated patterns is an NP-hard problem [3]. Therefore, this task constitutes the most consuming part, w.r.t. execution time, in RCPRMINER.

The next section experimentally studies the $\mathcal{RCPR}$ representation compactness.

5 Experimental Results

In this section, our main objective is to show, through extensive experiments, that the $\mathcal{RCPR}$ representation provides interesting compactness rates compared to the whole set of rare correlated patterns. All experiments were carried out on a PC equipped with a 2.7 GHz Intel Dual Core processor $E5400$ and 4 GB of main memory, running the Linux Ubuntu 10.04. The experiments were carried out on different dense and sparse benchmark datasets.[4] Representative results are plot by Figure 3.

According to the obtained experimental results, interesting reduction rates are obtained whether *minsupp* varies or *minbond* varies. The $\mathcal{RCPR}$ representation is indeed proved to be a perfect cover of the $\mathcal{RCP}$ set. In fact, the size of $\mathcal{RCPR}$ is always smaller than that of $\mathcal{RCP}$ set over the entire range of the support and *bond* thresholds. For example, considering the dense MUSHROOM dataset for *minsupp* = *35*% and *minbond* = *0.15*: $|\mathcal{RCPR}|$ = *1, 810*, while $|\mathcal{RCP}|$ = *100, 156*. In this situation, $\mathcal{RCPR}$ offers a reduction reaching approximately *98*%. These results are obtained thanks to the non-injectivity of the closure operator f_{bond} which gathers into disjoint subsets, *i.e.*, f_{bond} *equivalence classes*, patterns that have the same characteristics. This process avoids mining redundant patterns. Note that, in this case, $|\mathcal{MRCP}|$ = *1, 412* and $|\mathcal{CRCP}|$ = *652*. Since the $\mathcal{RCPR}$ representation corresponds to the union without redundancy of the $\mathcal{MRCP}$ and $\mathcal{CRCP}$, we always have $|\mathcal{RCPR}| \leq |\mathcal{MRCP}| + |\mathcal{CRCP}|$.

We present in the following the application of the proposed representation to the context of intrusion detection.

[4] Available at *http://fimi.cs.helsinki.fi/data*.

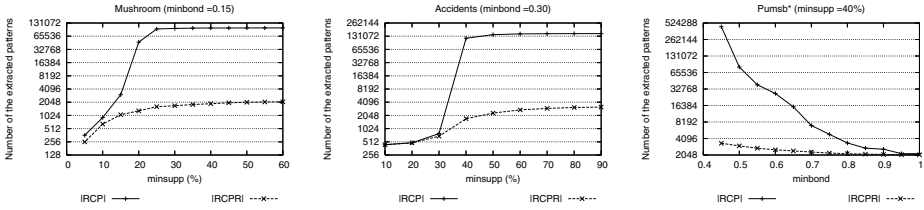

Fig. 3. Evaluation of the $\mathcal{RCPR}$ representation size w.r.t. the *minsupp* and *minbond* variations

6 Application to Intrusion Detection

We present in this section, the application of the $\mathcal{RCPR}$ representation in the design of an association rules based classifier. In fact, we used the $\mathcal{MRCP}$ and the $\mathcal{CRCP}$ sets, composing the $\mathcal{RCPR}$ representation, within the generation of the generic rare correlated rules of the form $Min \Rightarrow Closed \setminus Min$, with $Min \in \mathcal{MRCP}$ and $Closed \in \mathcal{CRCP}$.[5] Then, from the generated set of the generic rules, only the classification rules will be retained, *i.e.*, those having the label of the attack class in its conclusion part. After that, a dedicated classifier we designed is fed with these rules and has to perform the classification process and returns the detection rate for each attack class.

We present hereafter the application of our approach on the KDD 99 dataset.

6.1 Description of the KDD 99 Dataset

Each object of the KDD 99 dataset[6] represents a connection in the network data flow and is then labelled either normal or attack. KDD 99 defines 38 attacks categories partitioned into four `Attack` classes, which are DOS, PROBE, R2L and U2R, and one NORMAL class. The KDD 99 dataset contains 4, 940, 190 objects in the learning set and 41 input attributes for each connection. We propose in this work to consider 10% of the training set in the construction step of the classifier, containing 494, 019 objects. The learning set contains 79.20% (respectively, 0.83%, 19.65%, 0.22% and 0.10%) of DOS (respectively, PROBE, NORMAL, R2L and U2R).

6.2 Summary of Experimentations and Discussion of Obtained Results

Table 2 summarizes the obtained results, where AR and DR, respectively, denote "Association Rule" and "Detection Rate",[7] while *minconf* is the minimum threshold of the confidence measure [1]. In addition, by "Construction step", we mean that the step associated to the extraction of the $\mathcal{RCPR}$ representation while "Classification step"

[5] By "generic", it is meant that these rules are with minimal premises and maximal conclusions, w.r.t. set-inclusion.

[6] The KDD 99 dataset is available at the following link: *http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.*

[7] Detection Rate = $\frac{NbrCcCx}{TotalNbrCx}$, with $NbrCcCx$ stands for the number of the correctly classified connections and $TotalNbrCx$ is equal to the whole number of the classified connections.

represents the step in which the classification association rules are derived starting from $\mathcal{RCPR}$ and applied for detecting intrusions.

We note that the highest value of the detection rate is achieved for the classes NORMAL and DOS. In fact, this is related to the high number of connections of these two classes. This confirms that our proposed approach presents interesting performances even when applied to voluminous datasets. We also remark that the detection rate varies from an attack class to another one. In fact, for the U2R class, this rate is relatively low when compared to the others classes.

To sum up, according to Table 2, the computational cost varies from one attack class to another one. It is also worth noting that, for all the classes, the construction step is much more time-consuming than the classification step. This can be explained by the fact that the extraction of the $\mathcal{RCPR}$ concise representation is an NP-hard problem since the localization of the associated two borders is a complex task.

Table 2. Evaluation of the rare correlated association rules for the KDD 99 dataset

Attack class	***minsupp*** (%)	***minbond***	***minconf***	**# of generic exact ARs**	**# of generic approximate ARs**	**# of generic ARs of classification**	**CPU Time (in seconds) Construction step**	**Classification step**
DOS	**80**	**0.95**	**0.90**	**4**	**31**	**17**	**120**	**1**
PROBE	**60**	**0.70**	**0.90**	**232**	**561**	**15**	**55**	**1**
NORMAL	**85**	**0.95**	**0.95**	**0**	**10**	**3**	**393**	**15**
R2L	**80**	**0.90**	**0.70**	**2**	**368**	**1**	**1, 729**	**1**
U2R	**60**	**0.75**	**0.75**	**106**	**3**	**5**	**32**	**1**

Furthermore, the results shown by Table 3 prove that the proposed rare correlated association rules are more competitive than the decision trees as well as the Bayesian networks [2]. In fact, our approach presents better results for the attack classes DOS, R2L and U2R than these two approaches. For the NORMAL class, the obtained results using our approach are close to those obtained with the decision trees. The Bayesian networks based approach presents better detection rate only for the PROBE attack class. The proposed rare correlated association rules then constitute an efficient classification tool when applied to the intrusion detection in a computer network.

Table 3. Comparison between the proposed rare correlated association rules based classifier versus the state of the art approaches

Attack class	**Rare correlated generic ARs**	**Decision trees** [2]	**Bayesian networks** [2]
DOS	**98.68**	97.24	96.65
PROBE	70.69	77.92	**88.33**
NORMAL	**100.00**	99.50	97.68
R2L	**81.52**	0.52	8.66
U2R	**38.46**	13.60	11.84

7 Conclusion and Future Works

We proposed in this paper a characterization of the $\mathcal{RCP}$ set of rare correlated patterns and we defined the new exact concise $\mathcal{RCPR}$ representation associated with this set. We then designed the RCPRMINER algorithm allowing an efficient extraction of this representation. The carried out experimental studies highlight interesting compactness rates

offered by $\mathcal{RCPR}$. The effectiveness of the proposed classification method, based on generic rare correlated association rules, has also been proved in the context of intrusion detection. Other avenues of future works concern the extraction of generalized association rules starting from rare correlated patterns and their use in real-life applications. In addition, we plan to extend our approach to other correlation measures [6,10,12,14] through classifying them into classes of measures sharing the same properties.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), Santiago, Chile, pp. 487–499 (1994)
2. Ben Amor, N., Benferhat, S., Elouedi, Z.: Naive bayes vs decision trees in intrusion detection systems. In: Proceedings of the ACM Symposium on Applied Computing (SAC 2004), Nicosia, Cyprus, pp. 420–424 (2004)
3. Boley, M., Gärtner, T.: On the Complexity of Constraint-Based Theory Extraction. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds.) DS 2009. LNCS, vol. 5808, pp. 92–106. Springer, Heidelberg (2009)
4. Boulicaut, J.F., Jeudy, B.: Constraint-based data mining. In: Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 339–354. Springer (2010)
5. Ganter, B., Wille, R.: Formal Concept Analysis. Springer (1999)
6. Kim, S., Barsky, M., Han, J.: Efficient Mining of Top Correlated Patterns Based on Null-Invariant Measures. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part II. LNCS, vol. 6912, pp. 177–192. Springer, Heidelberg (2011)
7. Koh, Y.S., Rountree, N.: Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection. IGI Global Publisher (2010)
8. Kryszkiewicz, M.: Inferring Knowledge from Frequent Patterns. In: Bustard, D.W., Liu, W., Sterritt, R. (eds.) Soft-Ware 2002. LNCS, vol. 2311, pp. 247–262. Springer, Heidelberg (2002)
9. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 3(1), 241–258 (1997)
10. Omiecinski, E.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15(1), 57–69 (2003)
11. Pei, J., Han, J.: Constrained frequent pattern mining: a pattern-growth view. ACM-SIGKDD Explorations 4(1), 31–39 (2004)
12. Segond, M., Borgelt, C.: Item Set Mining Based on Cover Similarity. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS, vol. 6635, pp. 493–505. Springer, Heidelberg (2011)
13. Soulet, A., Crémilleux, B.: Adequate condensed representations of patterns. Data Mining and Knowledge Discovery 17(1), 94–110 (2008)
14. Surana, A., Kiran, R.U., Reddy, P.K.: Selecting a right interestingness measure for rare association rules. In: Proceedings of the 16th International Conference on Management of Data (COMAD 2010), Nagpur, India, pp. 115–124 (2010)
15. Taniar, D., Rahayu, W., Lee, V., Daly, O.: Exception rules in association rule mining. Applied Mathematics and Computation 205(2), 735–750 (2008)
16. Xiong, H., Tan, P.N., Kumar, V.: Hyperclique pattern discovery. Data Mining and Knowledge Discovery 13(2), 219–242 (2006)