Inferring Knowledge from Concise Representations of Both Frequent and Rare Jaccard Itemsets

Souad Bouasker[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. Correlated pattern mining has become increasingly an important task in data mining and knowledge discovery. Recently, concise exact representations dedicated for frequent correlated and for rare correlated patterns according to the *Jaccard* measure were presented. In this paper, we offer a new method of inferring new knowledge from the introduced concise representations. A new generic approach, called GMJP, allowing the extraction of the sets of frequent correlated patterns, of rare correlated patterns and their associated concise representations is introduced. Pieces of new knowledge in the form of associations rules can be either exact or approximate. We also illustrate the efficiency of our approach over several data sets and we prove that *Jaccard*-based classification rules have very encouraging results.

Keywords: Concise representation, Monotonicity, Constraint, Correlated pattern, *Jaccard* measure, Generic Approach.

1 Introduction and Motivations

Correlated item set mining is at the core of numerous data mining tasks. The enormous research efforts dedicated to this topic have led to a variety of sophisticated approaches [2,8,11,20,21]. In this regard, a variety of correlation measures were proposed and studied. In this work, we will focus on the *Jaccard* correlation measure [9]. Indeed, the *Jaccard* measure was used in many works under various names like *coherence* [14], *Tanimoto coefficient* [24] and *bond* measure [18] [(1)]. The *bond* measure was recently redefined in [3], where a concise exact representation of the set of frequent correlated patterns according to the *bond* measure was also proposed. Moreover, a generic approach for frequent *Jaccard* patterns mining was also performed in [20].

Frequent correlated itemset mining was then shown to be an interesting task in data mining. Since its inception, this key task grasped the interest of many researchers since it meets the needs of experts in several application fields [3], such as market basket study. However, the application of correlated frequent patterns is not an attractive solution for some other applications, *e.g.*, intrusion detection, analysis of the genetic confusion from biological data, pharmacovigilance, detection of rare diseases from medical data, to cite but a few [5,12,15,17,19,23]. As an illustration of the rare correlated patterns applications in the field of medicine, the rare combination of symptoms can provide useful

[1] In the rest of this paper, we used '*bond*' as a reference for the *Jaccard* measure.

H. Decker et al. (Eds.): DEXA 2013, Part II, LNCS 8056, pp. 109–123, 2013.

insights for doctors [25]. We cite that in [6], the authors proposed a concise exact representation of the set of rare correlated patterns and designed a rule-based classification process.

It is to mention, that no previous approach allowing the extraction of both frequent and rare correlated patterns according to a specified correlation metric was proposed. To solve this challenging problem, we propose an efficient algorithmic framework, called GMJP, allowing the extraction of both frequent correlated patterns, rare correlated patterns and their associated concise representations. To achieve the genericity of GMJP, we distinguish four different running scenarios depending on the required output. We, also, design a rule-based classifier and we discover meaningful correlations in data for frequent itemsets as well as for rare ones.

The paper is organized as follows. Section 2 presents the background used throughout this work. We introduce in Section 3 the recently proposed concise exact representations of both frequent and rare correlated patterns according to the *Jaccard* measure. Section 4 reviews some related work. The generic proposed approach GMJP is detailed in Section 5. We report an empirical study on different datasets and an association-rules based classifiction process respectively in Sections 6 and 7. We conclude and skecth issues of future work in Section 8.

2 Preliminaries

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:** $Supp(\vee I) = |\{t \in \mathcal{T} \mid (\exists\, i \in I, (t,i) \in \mathcal{R})\}|$

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 $Supp(\wedge \mathrm{AD}) = |\{1\}| = 1$ and $Supp(\vee \mathrm{AD}) = |\{1, 3, 5\}| = 3$.[2]

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

An itemset $\mathtt{I}$ is frequent if its support Supp($\mathtt{I}$) is above a user-defined minimum support threshold *minsupp*, otherwise the itemset $\mathtt{I}$ is said to be infrequent or rare.

The constraint of rarity is monotone, *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* [7] on the set of all the subsets of $\mathcal{I}$, $\mathcal{P}(\mathcal{I})$. Contrariwise, the frequency constraint induces an *order ideal* [7].

The *bond* measure [18] is mathematically equivalent to *Jaccard* [9]. It was redefined in [3] as:

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

The set of correlated patterns associated to the *bond* measure is defined as follows.

Definition 2. (***Correlated patterns***) *Considering a minimum correlation threshold minbond, the set $\mathcal{CP}$ of correlated patterns is equal to: $\mathcal{CP} = \{I \subseteq \mathcal{I} \mid bond(I) \geq minbond\}$* [3].

The *bond* measure takes its values within the interval $[0, 1]$. While considering the universe of a pattern I [14], *i.e.*, the set of transactions containing a non empty subset of I, the *bond* measure represents the simultaneous occurrence rate of the items of the pattern I in its universe. Thus, the more the items of I are dependent on each other, (*i.e.* strongly correlated), the higher the value of the *bond* measure is, since $Supp(\wedge I)$ would be closer to $Supp(\vee I)$. We present in what follows the concise exact representations associated to correlated patterns.

3 Concise Exact Representations of Correlated Patterns

In [3] and in [6], the authors introduced concise exact representations of respectively frequent correlated and rare correlated patterns. The proposed approaches are based on the concept of correlated equivalence classes induced by the f_{bond} closure operator associated to the *bond* measure.

In each equivalence class, all the elements have the same f_{bond} closure and the same value of *bond*. The minimal patterns of a *bond* equivalence class are the smallest incomparable members, w.r.t. set inclusion and are called **Minimal correlated patterns**, while the closed pattern is the largest one and is called **Closed correlated patterns**. These two sets are defined [3] as follows:

Definition 3. (***Closed correlated patterns by*** f_{bond}) *The set $\mathcal{CCP}$ of closed correlated patterns by f_{bond} is equal to: $\mathcal{CCP} = \{I \in \mathcal{CP} \mid \nexists\ I_1 \supset I$: $bond(I) = bond(I_1)\}$.*

Definition 4. (***Minimal correlated patterns***) *The set $\mathcal{MCP}$ of minimal correlated patterns is equal to: $\mathcal{MCP} = \{I \in \mathcal{CP} \mid \nexists\ I_1 \subset I$: $bond(I) = bond(I_1)\}$.*

While integrating the frequency constraint with the correlation constraint, we can distinguish between two sets of correlated patterns, which are the "Frequent correlated

[3] We refer in the rest of the paper to the minimum support threshold by *minsupp* and to the minimum correlation threshold by *minbond*.

patterns" set and the "Rare correlated patterns" set. Now, based on these two previous sets, the concise exact representation of frequent correlated patterns was studied and proposed in [3], in addition to the concise exact representation of rare correlated patterns which was proposed in [6].

Definition 5. (***The set of frequent correlated patterns***) *[3] The set $\mathcal{FCP}$ of frequent correlated patterns is equal to:* $\mathcal{FCP} = \{I \subseteq \mathcal{I} \mid Supp(\wedge I) \geq minsupp \text{ and } bond(I) \geq minbond\}$.

Definition 6. (***Concise exact representation of the $\mathcal{FCP}$ set***) *[3] The representation $\mathcal{RFCCP}$ is based on the set of frequent closed correlated patterns:*

$$\mathcal{RFCCP} = \{(I, Supp(\wedge I), Supp(\vee I)) \mid I \in \mathcal{CCP} \text{ and } Supp(\wedge I) >= minsupp\}.$$

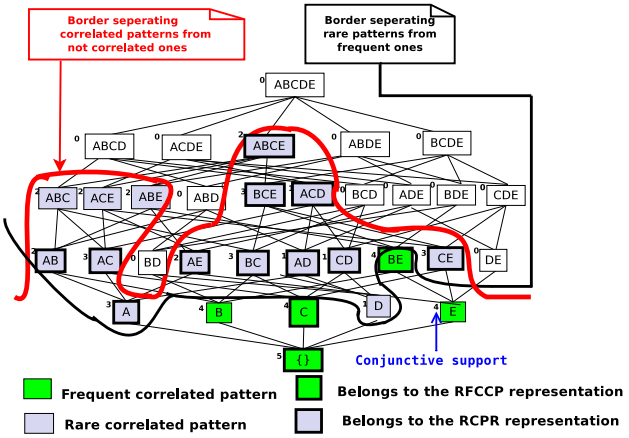

Fig. 1. Localization of the frequent correlated and the rare correlated patterns, and their associated condensed representation for *minsupp* = 4 and *minbond* = 0.2

Now, we present the rare correlated patterns associated to the *bond* measure:

Definition 7. (***The set of rare correlated patterns***) *[6] The $\mathcal{RCP}$ set of rare correlated patterns is equal to:* $\mathcal{RCP} = \{I \subseteq \mathcal{I} \mid Supp(\wedge I) < minsupp \text{ and } bond(I) \geq minbond\}$.

Definition 8. (***Concise exact representation of the $\mathcal{RCP}$ set***) *[6] Let $\mathcal{RCPR}$ be the concise exact representation of the $\mathcal{RCP}$ set based on the $\mathcal{CRCP}$ set of closed rare correlated patterns and on the $\mathcal{MRCP}$ set of the minimal rare correlated patterns. The $\mathcal{RCPR}$ representation is equal to:* $\mathcal{RCPR} = \mathcal{CRCP} \cup \mathcal{MRCP}$, *with*

$$\mathcal{CRCP} = \{(I, Supp(\wedge I), Supp(\vee I)) \mid I \in \mathcal{CCP} \text{ and } Supp(\wedge I) < minsupp\} \text{ and,}$$

$$\mathcal{MRCP} = \{(I, Supp(\wedge I), Supp(\vee I)) \mid I \in \mathcal{MCP} \text{ and } Supp(\wedge I) < minsupp\}.$$

These previous sets are depicted by Figure 1. The support shown at the top left of each frame represents the conjunctive support.

After presenting the *Jaccard* patterns, we propose in the next section, an overview of the approaches dealing with concise representations of correlated patterns.

4 Related Work

The problem of mining concise representations of correlated patterns was thoroughly studied in various works in the literature. The *bond* measure was studied in [13], the authors proposed an apriori-like algorithm for mining classification rules. Moreover, the authors in [20] proposed a generic approach for correlated patterns mining. Indeed, the *bond* correlation measure and eleven other correlation measures were used, all of them fulfill the anti-monotonicity property. Correlated patterns mining was then shown to be more complex and more informative than frequent patterns mining [20]. Also, in [22], a study of different properties of interesting measures was conducted in order to suggest a set of the most adequate properties to consider while mining rare associations rules. However, it is important to highlight that the extraction of rare correlated patterns was not carried out in [20] nor in [22].

Many other works have also emerged. In [26], the authors provide a unified definition of existing null-invariant correlation measures and propose the GAMINER approach allowing the extraction of frequent high correlated patterns according to the *Cosine* and to the *Kulczynski* measures. In this same context, the NICOMINER algorithm was also proposed in [10] and it allows the extraction of correlated patterns according to the *Cosine* measure. In this same context, we cite also the AETHERIS approach [21] which allow the extraction of condensed representation of correlated patterns according to user's preferences. In [2], the authors introduced the concept of flipping correlation patterns according to the *Kulczynsky* measure. However, the *Kulczynsky* does not fulfill the interesting anti-monotonic property as the *bond* measure. To the best of our knowledge, this work is the first one that puts the focus on mining concise representations of both frequent and rare correlated patterns according to the *bond* measure.

We introduce, in what follows, our new GMJP approach [4].

5 The GMJP Approach

We introduce in this section the GMJP approach which allows, according to the user's input parameters, the extraction of the desired output. As shown by Figure 2, four different scenarios are possible for running the GMJP approach:

- **First Scenario:** outputs the whole set $\mathcal{FCP}$ of frequent correlated patterns,
- **Second Scenario:** outputs the $\mathcal{RFCCP}$ concise exact representation of the $\mathcal{FCP}$ set,
- **Third Scenario:** outputs the whole set $\mathcal{RCP}$ of rare correlated patterns,
- **Fourth Scenario:** outputs the $\mathcal{RCPR}$ concise exact representation of the $\mathcal{RCP}$ set.

The GMJP algorithm takes as an input a dataset $\mathcal{D}$, a minimal support threshold *minsupp* and a minimal correlation threshold *minbond*. We mention that GMJP determines exactly the *support* and the *bond* values of each pattern of the desired output according to the user's parameters.

[4] GMJP stands for **G**eneric **M**ining of **J**accard **P**atterns.

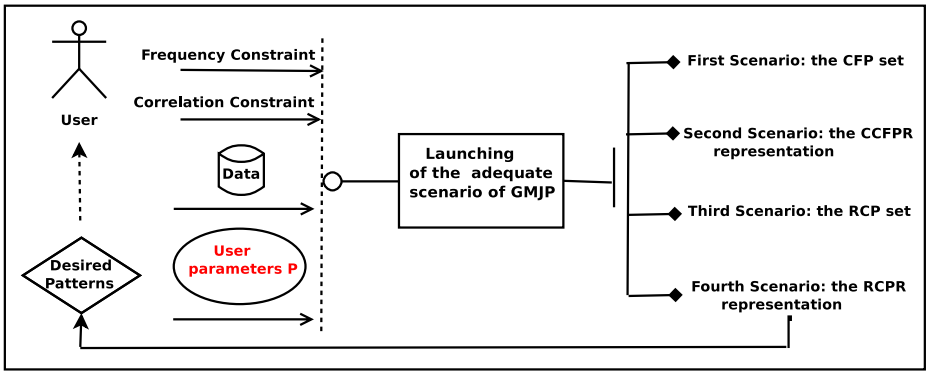

Fig. 2. Overview of GMJP

5.1 Overview of the Algorithm

We illustrate the different steps of GMJP when running the fourth script aiming to extract the $\mathcal{RCPR}$ representation. Our choice of this fourth scenario is motivated by the fact that the extraction of the $\mathcal{RCPR}$ representation corresponds to the most challenging mining task for GMJP.

In fact, $\mathcal{RCPR}$ is composed by the set of rare correlated patterns which results from the intersection of two theories [16] induced by the constraints of correlation and rarity. So, this set is neither an order ideal nor an order filter. Therefore, the localization of the elements of the $\mathcal{RCPR}$ representation is more difficult than the localization of theories corresponding to constraints of the same nature. Indeed, the conjunction of anti-monotonic constraints (*resp.* monotonic) is an anti-monotonic constraint (*resp.* monotonic) [4]. For example, the constraint "being a correlated frequent pattern" is anti-monotonic, since it results from the conjunction of two anti-monotonic constraints namely, "being a correlated pattern" and "being a frequent pattern". This constraint induces, then, an order ideal on the itemsets lattice. In fact, the GMJP algorithm mainly operates in three steps as depicted by Figure 3.

1. A first scan of the dataset is performed in order to extract all the items and assigning to each item the set of transactions in which it appears. Then, a second scan of the dataset is carried out in order to identify, for each item, the list of the co-occurrent items.
2. The second step consists in integrating both the constraints of rarity and of correlation in a mining process of $\mathcal{RCPR}$. In this situation, this problem is split into independent chunks since each item is treated separately. In fact, for each item, a set of candidates is generated. Once obtained, these candidates are pruned using the following pruning strategies:

   (***a***) **The pruning of the candidates which check the cross-support property [3].**
   (***b***) **The pruning based on the order ideal of the correlated patterns**.

   Recall that the set of correlated patterns induces an order ideal property. Therefore, each correlated candidate, having a non correlated subset, will be pruned since it will not be a correlated pattern. Then, the conjunctive, disjunctive supports and the

bond value of the retained candidates are computed. Thus, the uncorrelated candidates are also pruned. At the level n, the local minimal rare correlated patterns of size n are determined among the retained candidates. The local closed rare correlated patterns of size $n-1$ are also filtered. This process holds up when there is no more candidates to be generated.

3. The third and last step consists of filtering the global minimal rare correlated patterns and the global rare correlated patterns among the two sets of local minimal rare correlated patterns and local closed ones.

In what follows, we will explain more deeply these different steps of GMJP. The pseudo code is given by Procedure 1.

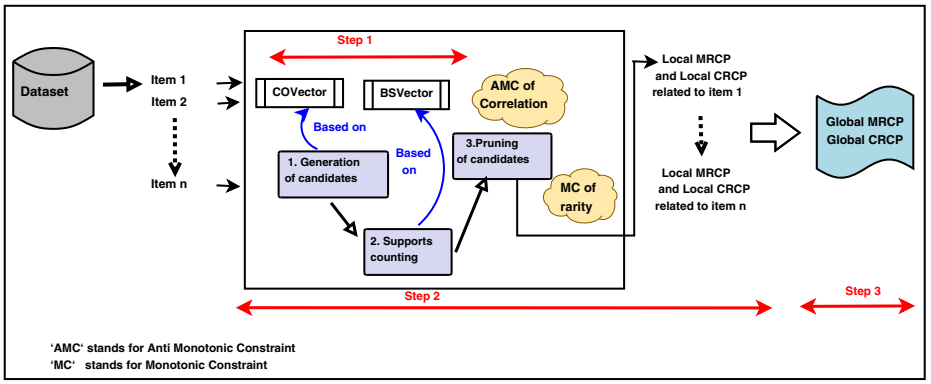

Fig. 3. Overview of GMJP when extracting the $\mathcal{RCPR}$ representation

First Step: The Power of the Bit Vectors and of Co-occurrent Vectors. Initially, the dataset is scanned in order to extract the items and to build, for each item, the bitset called here "BSVector". In fact, a bitset is a container that can store a huge number of bits while optimizing the memory consumption (For example, 32 elements are stored in a memory block of 4 bytes). Each block of memory is treated in just one CPU operation by a 32 bits processor. Therefore, we are very motivated for these kinds of structures within the GMJP algorithm in order to optimize the conjunctive and the disjunctive supports computations.

Then, the dataset is scanned again in order to identify, for each item I, the list of the co-occurrent items which corresponds to the items occurring in the same transactions as the item I. These latter ones are stored in a vector of integers, called here "COVector". We note that one of the main challenges of the GMJP algorithm is that it allows pushing two constraints of distinct types and to deliver the output with only two scans of the dataset. We uphold also that the bitsets, when incorporated into the mining process within the GMJP algorithm, sharply decrease the size of the memory required to store immediate results and significantly save execution costs.

Second Step: Getting the Local Minimal and the Local Closed Rare Correlated Patterns without Closure Computations. Worth of mention, the main thrust of the GMJP algorithm is to break the search space into independent sub-problems. In fact, for

Algorithm 1. GMJP

Data:
1. A dataset $\mathcal{D}$.
2. A minimal correlation threshold *minbond*.
3. A minimal conjunctive support threshold *minsupp*.
4. A specification of the desired result '$\mathcal{RCPR}$'.

Results: The concise exact representation $\mathcal{RCPR} = \mathcal{MRCP} \cup \mathcal{CRCP}$.

Begin

1. Scan the dataset $\mathcal{D}$ twice to build the BSVector and the COVector for all the items
2. For each item I
   (a) n =2;
   (b) Generate the candidates of size n using the COVector of I
   (c) **While** (*The number of the generated candidates is not null*) **Do**
      i. Prune these candidates w.r.t. the cross-support property of the *bond* measure
      ii. Prune these candidates w.r.t. the order ideal property of correlated patterns
      iii. Compute the conjunctive and disjunctive supports and the *bond* value of the maintained candidates
      iv. For each candidate C
         If (IsCorrelated(C) and IsRare(C)) **then**
         $/*$ Ckeck-Local Minimality of the candidate C $*/$
            – Update the set of Local Minimal Rare Correlated Patterns of size n
      v. Find Local Closed Rare Correlated Patterns of size $n-1$
      vi. n = $n+1$
      vii. Generate candidates of size n using the APRIORI-GEN procedure
3. Find all Global Minimal Rare Correlated Patterns
4. Find all Global Closed Rare Correlated Patterns
5. **Return** $\mathcal{RCPR}$;

End

each item I, a levelwise mining process is performed using the COVector containing the co-occurrent items of I. At each level n, starting by the second level, a set of candidates are generated, then pruned according to the different pruning strategies described previously. The minimal rare correlated patterns of size n associated to the item I are called **Local Minimal Rare Correlated Patterns** and they are determined by comparing their *bond* values to those of their respective immediate subsets. Similarly, the closed rare correlated patterns of size $n - 1$ associated to the item I are called **Local Closed Rare Correlated Patterns** and they are determined by comparing their *bond* values to those of their respective immediate supersets.

It is also important to mention that the implementation of the different stages of this second step (candidate generation, evaluation and pruning) was based on simple vectors of integers. Thus, we require no more complex data structure during the implementation of the GMJP algorithm. This feature makes GMJP a practical approach for handling both monotonic and anti-monotonic constraints even for large datasets.

One of the major challenges in the design of the GMJP algorithm is how to perform subset and superset checking to efficiently identify Local Minimal and Local Closed patterns? The answer is to construct and manage a multimap hash structure, [5] in order to store at each level n the rare correlated patterns of size n. This technique is very powerful since it makes the subset and the superset checking practical even on dense datasets.

Thus, our proposed efficient solution (as we prove it experimentally later) is to integrate both the monotonic constraint of rarity and the anti-monotonic constraint of correlation into the mining process and to identify the local closed rare correlated patterns without closure computing.

Third Step: Filtering the Global Minimal and the Global Closed Rare Correlated Patterns. After identifying the local minimal and the local closed rare correlated patterns associated to each item I of the dataset $\mathcal{D}$, the third step consists in filtering the $\mathcal{MRCP}$ set of Global Minimal Rare Correlated patterns and the $\mathcal{CRCP}$ set of Global Closed Rare Correlated patterns. This task is performed using two distinct multimap hash structures. In fact, for each local minimal rare correlated pattern LM previously identified, we check whether it has a direct subset (belonging to the whole set of local minimal patterns) with the same *bond* value. If it is not the case, then the local minimal pattern LM is a global minimal rare pattern and it is added to the $\mathcal{MRCP}$ set. Similarly, for each local closed rare correlated pattern LC previously identified, we check whether it has a direct superset (belonging to the whole set of local closed patterns) with the same *bond* value. If it is not the case, then the local closed pattern LC is a global closed rare pattern and it is added to the $\mathcal{CRCP}$ set of Closed rare correlated patterns.

In what follows, we illustrate with a running example of the GMJP algorithm.

5.2 A Running Example

Let us consider the dataset $\mathcal{D}$ given by Table 1. First, the BSVectors and the COVectors associated to each item of this dataset are constructed, as we plot by Figure 4. These BSVectors are next used to compute the conjunctive and the disjunctive supports. We have, for example, the item A which belongs to the transactions $\{1, 3, 5\}$ and the item C which belongs to the transactions $\{1, 2, 3, 5\}$. We, then, have $Supp(\wedge AC) = 3$ and $Supp(\vee AC)) = 4$.

The local minimal and the local closed correlated rare patterns associated to each item I of the dataset $\mathcal{D}$, are extracted. A detailed example of the process of the item A is given by Figure 5. The finally obtained $\mathcal{RCPR}$ representation, for *minsupp* = 4 and for *minbond* = 0.20, is composed by the following global minimal and global closed

[5] We used in our implementation the C++ STL Standard Template Library multimap.

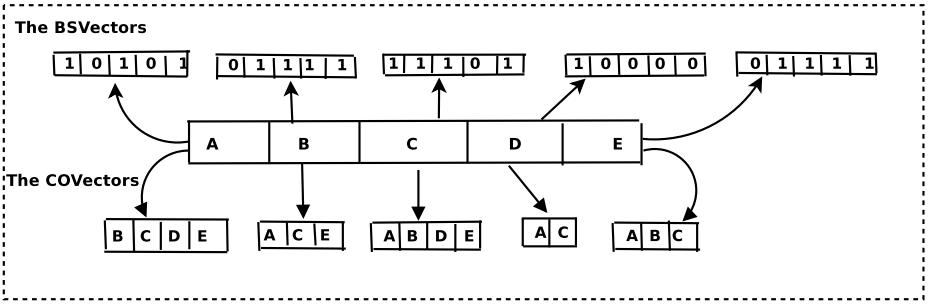

Fig. 4. The BSVectors and the COVectors associated to the items of the dataset $\mathcal{D}$

Jaccard patterns: $\mathcal{RCPR}$ = { $(A, 3, \frac{3}{3})$, $(D, 1, \frac{1}{1})$, $(AB, 2, \frac{2}{5})$, $(AC, 3, \frac{3}{4})$, $(AD, 1, \frac{1}{3})$, $(AE, 2, \frac{2}{5})$, $(BC, 3, \frac{3}{5})$, $(CD, 1, \frac{1}{4})$, $(CE, 3, \frac{3}{5})$, $(ACD, 1, \frac{1}{4})$, $(BCE, 3, \frac{3}{5})$ and $(ABCE, 2, \frac{2}{5})$}.

Last, it is important to notice that GMJP is not an exclusive approach in the sense that it can be coupled with other efficient approaches to mine statistically significant patterns.

In the next section we report our experimental study of the proposed GMJP algorithm.

6 Experimental Evaluation

Datasets and Experimental Environment: Experiments were carried out on different dense and sparse benchmark datasets [6]. All the tests were carried out on a PC equipped with a 2.40 GHz Intel Core TM $i3$ processor and 2.92 GB of main memory, running the Linux Ubuntu 10.04. Running times were averaged over 5 executions.

Protocol: Our objective is to prove, through extensive carried out experiments, the efficiency of the proposed GMJP algorithm while running the four different scenarios. Our first batch of experiments aims to build a quantitative comparison between the $\mathcal{FCP}$, the $\mathcal{RCP}$ sets and their associated condensed representations. Our second batch of experiments focus on studying running times.

Results: As sketched by Table 2, the concise representations $\mathcal{FCCPR}$ and $\mathcal{RCPR}$ present very encouraging reduction rates over several datasets and for different ranges of *minsupp* and *minbond* theresholds. We note that, the 'gain' corresponds to the reduction rate and is equal to : 1 - $\frac{|\mathcal{RCPR}|}{|\mathcal{RCP}|}$ for rare *Jaccard* patterns, and equal to 1 - $\frac{|\mathcal{FCCPR}|}{|\mathcal{FCP}|}$ for frequent ones.

[6] Available at *http://fimi.cs.helsinki.fi/data* and at http://archive.ics.uci.edu/ml.

We conclude, according to the results gived by Table 3 [7] , that the execution time varies depending on the number of distinct items of the considered dataset. This is explained by the principle of GMJP which is based on the idea of processing each item separately and based on the list of the co-occurrent of each item. For example, the computational costs are relatively high for the T40I10D100K dataset, and they are lower for the MUSHROOM dataset. This is explained by the fact that, the MUSHROOM dataset contains only *119* items while the T40I10D100K dataset contains *942* items. We note also that the highest execution times are obtained with the RETAIL dataset, since this latter contains a high number of distinct items, equal to 16,470.

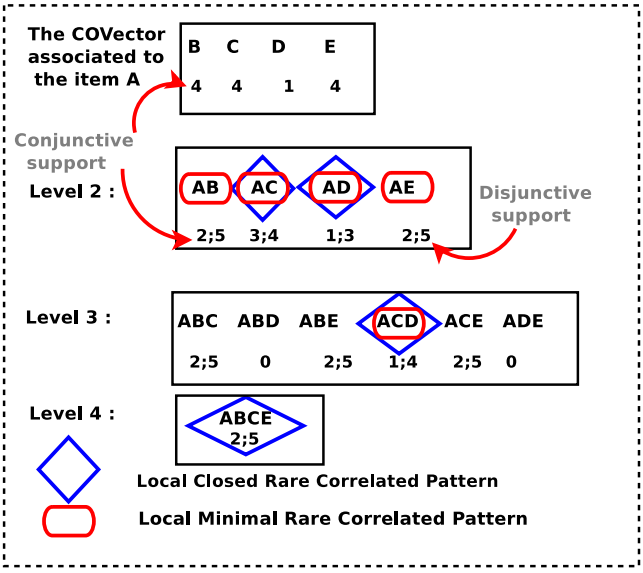

Fig. 5. Mining Local Minimal and Local Closed Rare Correlated Patterns for the item A

It is worth of mention that the computational time of the fourth scenario dedicated to the extraction of the $\mathcal{RCPR}$ are the highest ones. This can be explained by the fact that the extraction of the $\mathcal{RCPR}$ representation is an NP-hard problem since the localization of the associated two borders is a complex task. We also highlight that the performance results of the designed GMJP algorithm can not be compared to any approach of the literature. Indeed, the proposed approach is the first one dedicated to the extraction of *Jaccard* patterns in a generic way.

In the next section, we study the process of classification based on correlated association rules derived from the previous presented condensed representations.

[7] We note that 'S1' stands for the First Scenario, 'S2' stands for the Second Scenario, 'S3' stands for the Third Scenario and 'S4' stands for the Fourth Scenario.

Table 2. Effectiveness of *Jaccard* patterns mining on UCI benchmarks

Dataset	*minsupp*	*minbond*	# $\mathcal{FCP}$	# $\mathcal{FCCPR}$	Gain of $\mathcal{FCCPR}$	# $\mathcal{RCP}$	# $\mathcal{RCPR}$	Gain of $\mathcal{RCPR}$
MUSHROOM	30%	0.15	2, 701	427	**84.19**%	98, 566	1, 704	**98.27**%
	45%	0.15	307	83	**72.96**%	100, 960	1, 985	**98.03**%
PUMSB*	40%	0.45	10, 674	1646	**84.57**%	448, 318	3, 353	**99.25**%
	40%	0.50	9, 760	1325	**86.42**%	82, 413	3, 012	**96.34**%
CONNECT	10%	0.80	534, 026	15, 152	**97.16**%	56	56	**0**%
	50%	0.80	533, 991	15, 117	**97.16**%	91	91	**0**%
ACCIDENTS	40%	0.30	32, 529	32, 528	**0**%	117, 805	1, 722	**98.53**%
	60%	0.30	2, 057	2, 047	**0**%	148, 259	2, 743	**98.14**%

Table 3. Performance Analysis of GMJP on UCI benchmarks (time in second)

Dataset	Number of Items	Average *minsupp*	Average *minbond*	Average Time S1	Average Time S2	Average Time S3	Average Time S4
MUSHROOM	119	58%	0.30	7	11.4	20	19.6
		40%	0.57	3.75	5.25	11	709
ACCIDENTS	468	7.8%	0.50	709	703	793	784.2
RETAIL	16, 470	25.83%	0.50	5.83	13.16	1903	1902
T10I4D100K	870	5%	0.20	2	3	163	163
T40I10D100K	942	8.2%	0.50	148	182.6	491	490.4

7 Association Rules-Based Classification Process

We present in this section, the application of the $\mathcal{RCPR}$ and the $\mathcal{RFCCP}$ representations 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 [8] rare correlated rules. The $\mathcal{RFCCP}$ representation is used to generate generic frequent correlated rules, of the form $Min \Rightarrow Closed \setminus Min$, with Min is a minimal generator and $Closed$ is a closed pattern. Hence, we implemented a C++ program allowing the extraction of the correlated frequent minimal generators. Then, from the generated set of the generic rules, only the classification rules will be retained, *i.e.*, those having the label of the 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 accuracy rate for each class.

We report in Table 4 [9] the impact of integrating the correlation constraint for a fixed *minsupp* and *minconf* thresholds. We remark, for the frequent patterns, that while

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

[9] We note that Accuracy Rate = $\frac{NbrCcTr}{TotalNbrTr}$, with $NbrCcTr$ stands for the number of the correctly classified transactions and $TotalNbrTr$ is equal to the whole number of the classified transactions, and *minconf* corresponds to the minimum threshold of the confidence measure [1].

Table 4. Evaluation of the classification accuracy *versus minbond* variation for frequent and rare *Jaccard* patterns

Dataset	*minsupp*	*minconf*	*minbond*	# Exact Rules	# Approximate Rules	# Classification Rules	Accuracy rate	Response Time (sec)	Property of Patterns
WINE	1%	0.60	0	387	5762	650	97.75%	1000	Frequent
			0.10	154	2739	340	95.50%	13.02	Frequent
			0.20	60	1121	125	94.38%	1.00	Frequent
			0.30	20	319	44	87.07%	0.01	Frequent
ZOO	50%	0.70	0.30	486	2930	235	89.10%	40	Rare
			0.40	149	436	45	89.10%	3	Rare
			0.50	38	88	11	83.16%	0.01	Rare
			0.60	12	31	6	73.26%	0.01	Rare

increasing the *minbond* threshold, the number of exact and approximate association rules decreases while maintaining always an important accuracy rate. Another benefit for *Jaccard* measure integration, is the improvement of the response time, it varies from *1000* to *0.01* seconds. Whereas, for the rare patterns, we highlight that the increase of the *minbond* threshold induces a reduction in the accuracy rate. This is explained by the decrease in the number of the obtained classification rules.

Table 5. Evaluation of the classification accuracy of frequent patterns *vs* rare patterns

Dataset	*minbond*	*minsupp*	*minconf*	# Exact Rules	# Approximate Rules	# Classification Rules	Accuracy rate	Property of *Jaccard* patterns
WINE	0.1	20%	0.60	7	274	25	76.40%	Frequent
			0.80	7	86	10	86.65%	Frequent
			0.90	7	30	4	84.83%	Frequent
	0.1	20%	0.60	91	1516	168	**95.50%**	**Rare**
			0.80	91	449	84	92.69%	Rare
			0.90	91	100	48	91.57%	Rare
IRIS	0.15	20%	0.60	3	22	7	**96.00%**	**Frequent**
			0.95	3	6	3	95.33%	Frequent
	0.15	20%	0.60	17	32	8	80.06%	Rare
			0.95	17	7	5	80.00%	Rare
	0.30	20%	0.60	3	22	7	**96.00%**	**Frequent**
			0.95	3	6	3	95.33%	Frequent
	0.30	20%	0.60	8	14	4	70.00%	Rare
			0.95	8	6	3	69.33%	Rare
TICTACTOE	0	10%	0.80	0	16	16	**69.40%**	Frequent
	0.05	10%	0.80	0	16	16	**69.40%**	Frequent
	0.07	10%	0.80	0	8	8	**63.25%**	Frequent
	0.1	10%	0.80	0	1	1	**60.22%**	Frequent
	0	10%	0.80	1, 033	697	192	**100.00%**	**Rare**
	0.05	10%	0.80	20	102	115	**100.00%**	**Rare**
	0.07	10%	0.80	8	66	69	**97.07%**	Rare
	0.1	10%	0.80	2	0	1	**65.34%**	Rare

We note according to the results sketched by Table 5, that for the datasets WINE and TICTACTOE , the highest values of the accuracy rate are achieved with the rare correlated rules. Whereas, for the IRIS dataset, the frequent correlated rules performed higher accuracy than rare ones. In this regard, we can conclude that for some datasets,

the frequent correlated patterns have better informativity than rare ones. Whereas, for other datasets, rare correlated patterns bring more rich knowledge. This confirms the beneficial contribution of our approach in inferring new knowledge from both frequent and rare *Jaccard* patterns.

8 Conclusion and Future Works

We proposed, in this paper, GMJP the first approach to mine *Jaccard* patterns in a generic way (i.e., with two types of constraints: anti-monotonic constraint of frequency and monotonic constraint of rarity). Our approach is based on the key notion of bit-sets codification that supports efficient *Jaccard* patterns computation thanks to an adequate condensed representation of patterns. Experiments realised on several datasets show the efficiency of GMJP according to both quantitative and qualitative aspects. An important direction for future work is to extend our approach to other correlation measures [10,18,20,22] 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. Barsky, M., Kim, S., Weninger, T., Han, J.: Mining flipping correlations from large datasets with taxonomies. In: Proceedings of the 38th International Conference on Very Large Databases, VLDB 2012, Istanbul, Turkey, pp. 370–381 (2012)
3. Ben Younes, N., Hamrouni, T., Ben Yahia, S.: Bridging conjunctive and disjunctive search spaces for mining a new concise and exact representation of correlated patterns. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 189–204. Springer, Heidelberg (2010)
4. Bonchi, F., Lucchese, C.: On condensed representations of constrained frequent patterns. Knowledge and Information Systems 9(2), 180–201 (2006)
5. Booker, Q.E.: Improving identity resolution in criminal justice data: An application of NORA and SUDA. Journal of Information Assurance and Security 4, 403–411 (2009)
6. Bouasker, S., Hamrouni, T., Ben Yahia, S.: New exact concise representation of rare correlated patterns: Application to intrusion detection. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012, Part II. LNCS, vol. 7302, pp. 61–72. Springer, Heidelberg (2012)
7. Ganter, B., Wille, R.: Formal Concept Analysis. Springer (1999)
8. Grahne, G., Lakshmanan, L.V.S., Wang, X.: Efficient mining of constrained correlated sets. In: Proceedings of the 16th International Conference on Data Engineering (ICDE 2000), pp. 512–521. IEEE Computer Society Press, San Diego (2000)
9. Jaccard, P.: Étude comparative de la distribution orale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901)
10. 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)
11. Kim, W.-Y., Lee, Y.-K., Han, J.: CCMine: Efficient mining of confidence-closed correlated patterns. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 569–579. Springer, Heidelberg (2004)

12. Koh, Y.S., Rountree, N.: Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection. IGI Global Publisher (2010)
13. Le Bras, Y., Lenca, P., Lallich, S.: Mining classification rules without support: an anti-monotone property of jaccard measure. In: Elomaa, T., Hollmén, J., Mannila, H. (eds.) DS 2011. LNCS, vol. 6926, pp. 179–193. Springer, Heidelberg (2011)
14. Lee, Y.K., Kim, W.Y., Cai, Y.D., Han, J.: CoMine: efficient mining of correlated patterns. In: Proceedings of the 3rd International Conference on Data Mining (ICDM 2003), pp. 581–584. IEEE Computer Society Press, Melbourne (2003)
15. Mahmood, A.N., Hu, J., Tari, Z., Leckie, C.: Critical infrastructure protection: Resource efficient sampling to improve detection of less frequent patterns in network traffic. Journal of Network and Computer Applications 33(4), 491–502 (2010)
16. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 3(1), 241–258 (1997)
17. Manning, A.M., Haglin, D.J., Keane, J.A.: A recursive search algorithm for statistical disclosure assessment. Data Mining and Knowledge Discovery 16(2), 165–196 (2008)
18. Omiecinski, E.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15(1), 57–69 (2003)
19. Romero, C., Romero, J.R., Luna, J.M., Ventura, S.: Mining rare association rules from e-learning data. In: Proceedings of the 3rd International Conference on Educational Data Mining (EDM 2010), Pittsburgh, PA, USA, pp. 171–180 (2010)
20. 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)
21. Soulet, A., Raissi, C., Plantevit, M., Crémilleux, B.: Mining dominant patterns in the sky. In: Proceedings of the 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, Canada, pp. 655–664 (2011)
22. 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)
23. Szathmary, L., Valtchev, P., Napoli, A.: Generating rare association rules using the minimal rare itemsets family. International Journal of Software and Informatics 4(3), 219–238 (2010)
24. Tanimoto, T.T.: An elementary mathematical theory of classification and prediction. Technical Report, I.B.M. Corporation Report (1958)
25. Tsang, S., Koh, Y.S., Dobbie, G.: RP-tree: Rare pattern tree mining. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2011. LNCS, vol. 6862, pp. 277–288. Springer, Heidelberg (2011)
26. Wu, T., Chen, Y., Han, J.: Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining and Knowledge Discovery 21, 371–397 (2010)