Generating Rare Association Rules Using the Minimal Rare Itemsets Family

Laszlo Szathmary, Petko Valtchev, Amedeo Napoli

Int J Software Informatics, Vol.4, No.3, September 2010, pp. 219–238 E-mail: ijsi@iscas.ac.cn
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org
 Tel: +86-10-62661040

Generating Rare Association Rules Using the Minimal Rare Itemsets Family

Laszlo Szathmary[1], Petko Valtchev[1], and Amedeo Napoli[2]

[1] (Dépt. d'Informatique UQAM, C.P. 8888, Succ. Centre-Ville, Montréal H3C 3P8, Canada)

[2] (LORIA UMR 7503, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France)

Abstract Rare association rules correspond to rare, or infrequent, itemsets, as opposed to frequent ones that are targeted by conventional pattern miners. Rare rules reflect regularities of local, rather than global, scope that can nevertheless provide valuable insights to an expert, especially in areas such as genetics and medical diagnosis where some specific deviations/illnesses occur only in a small number of cases. The work presented here is motivated by the long-standing open question of efficiently mining strong rare rules, i.e., rules with high confidence and low support. We also propose an efficient solution for finding the set of minimal rare itemsets. This set serves as a basis for generating rare association rules.

Key words: data mining; knowledge discovery in databases (KDD); itemset extraction; rare itemsets; rare association rules; rare item problem

Szathmary L, Valtchev P, Napoli A. Generating rare association rules using the minimal rare itemsets family. *Int J Software Informatics*, **2010, 4(3): 219–238. http://www.ijsi.org/1673-7288/4/i56.htm**

1 Introduction

Conventional pattern miners target the frequent itemsets and rules in a dataset. These are believed to reflect the globally valid trends and regularities dug in the data, hence they typically support modelling and/or prediction. Yet in many cases global trends are known or predictable beforehand by domain experts, therefore such patterns do not bear much value to them. In contrast, regularities of local scope, i.e., covering only a small number of data records, or transactions, may be of higher interest as they could translate less well-known phenomena, e.g., contradictions to the general beliefs in the domain or notable exceptions thereof [16]. This is often true in areas such as genetics and medical diagnosis where many deviations / symptom combinations will only manifest in a small number of patient cases. Hence the potential of the methods for mining the corresponding patterns and rules for supporting a more focused analysis of the recorded biomedical data. The present paper is a revised and extended version of [23] and [24].

1.1 Motivating examples

A first case study for atypical patterns and rules pertains to a French biomedical database, the STANISLAS cohort [17]. The STANISLAS cohort comprises the medical

Corresponding author: Laszlo Szathmary; e-mail: Szathmary.L@gmail.com
Received 2010-08-18; revised 2010-10-01; accepted 2010-10-14

records of a thousand presumably healthy French families. In a particular problem settings, the medical experts are interested in characteristics and relations that pertain to a very small number of individuals. For instance, a key goal in this context is to investigate the impact of genetic and environmental factors on diversity in cardiovascular risk factors. Interesting information to extract from the cohort database includes the patient profiles associating genetic data with extreme or borderline values of biological parameters. However, such types of associations should be atypical in healthy cohorts.

To illustrate the concept of rare rules and its potential benefits, assume we want to target the causes for a group of cardiovascular diseases (CVD) within the STANISLAS cohort. If a frequent combination of CVD and a potential factor is found, then the factor may be reasonably qualified as a facilitator for the disease. For instance, a frequent itemset "{elevated cholesterol level, CVD}" and a strong association rule "{elevated cholesterol level}⇒ {CVD}" would empirically validate the widely acknowledged hypothesis that people with high cholesterol level are at serious risk of developing a CVD. In contrast, if the itemset involving a factor and CVD is rare, this would suggest an inhibiting effect on the disease. For instance, the rareness of the itemset "{vegetarian, CVD}" would suggest that a good way to reduce the CVD risk is to observe a vegetarian diet.

The second case study pertains to pharmacovigilance, a domain of pharmacology dedicated to the detection, monitoring and study of adverse drug effects. Given a database of clinical records together with taken drugs and adverse effects, mining relevant itemsets would enable a formal association between drugs adverse effects. Thus, the detected patterns of (combinations of) drugs with undesired (or even fatal) effects on patients could provide the basis for an informed decision as to the withdrawal or continuance of a given drug. Such decision may affect specific patients, part of or even in the entire drug market (see, for instance, the withdrawal of the lipid-lowering drug *Cerivastatin* in August 2001). Yet in order to make appear the alarming patterns of adverse effects, the benign ones, which compose the bulk of the database content, should be filtered out first. Once again, there is a need to skip the typical phenomena and to focus on less expectable ones. It is noteworthy that similar reasoning may be abstracted from unrelated problem domains such as bank fraud detection where fraudulent behaviour patterns manifest in only a tiny portion of the transaction database content.

1.2 State of the art

Pattern mining based on the support metrics is biased upon the detection of trends that are – up to a tolerance threshold – globally valid. Hence a straightforward approach to the detection of atypical and local regularities has been to relax the crisp and uniform minimal support criterion for patterns [26].

In a naïve problem settings, the minimal support could be decreased sufficiently to include in the frequent part of the pattern family all potentially interesting regularities. Yet this would have a devastating impact on the performances of the pattern miner on top of the additional difficulties in spotting the really interesting patterns within the resulting huge output (known as the *rare item problem* [15, 29]).

A less uniform support criterion is designed in [29] where the proposed method *RSAA* (Relative Support Apriori Algorithm) relies on item-wise minimal support thresholds with user-provided values. *RSAA* outputs all itemsets, and hence rules, having their support above at least one support threshold corresponding to a member item. Thus, the output still comprises all frequent itemsets and rules together with some, but not necessarily all, atypical ones.

A higher degree of automation is achieved in *MSapriori* (Multiple Supports Apriori) [15] by modulating the support of an itemset with the supports of its member items. Thus, the support is increased by a factor inversely proportional to the lowest member support, which, on the bottom line increases the chances of itemsets involving infrequent items to nevertheless make it to the frequent part of the pattern family. Once more, the overall effect is the extension of the frequent part in the pattern family by some infrequent itemsets.

In [28], Wu *et al.* proposes an extension of the traditional association rule mining framework to include rules of forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$, which indicate *negative associations* between itemsets. Negative association rules are obtained using infrequent itemsets. In contrast to positive association rules, negative association rules provide information about the absence of certain itemsets. *Emerging patterns* are itemsets whose support increases significantly from one dataset to another. Emerging patterns are said to capture emerging trends in time-stamped databases, or to capture differentiating characteristics between classes of data. Emerging patterns can have low support in dataset D_1 and high support in D_2, thus they can yield some important changes between the two datasets. See [19] for a survey on emerging patterns. In [20], the authors are interested in the extraction of concepts with smaller support in a given lattice. This work is carried out in the framework of Formal Concept Analysis [8] and is related to our work. However, our search for rare itemsets and rare association rules (with high confidence) is directly performed on data rather than exploring concepts within a concept lattice.

Our own approach is a more radical departure from the standard pattern mining settings as it focuses directly on the infrequent part of the pattern family that becomes the mining target. The underlying key notion is the *rare itemset (rule)* defined as an itemset (rule) with support lower than the threshold. *Apriori-Inverse*[10], and *MIISR* (Mining Interesting Imperfectly Sporadic Rules) [11] are two methods from the literature that exploit the same rarity notion, yet the former would exclusively mine perfectly rare itemsets (i.e., having exclusively rare subsets) while the latter slightly relaxes this overtly crisp constraint. This, on the bottom line, amounts to exploring rare patterns within the order filter above the rare singleton itemsets (i.e., rare items) in the itemset lattice while ignoring rare itemsets mixing both rare and frequent items.

Here we propose a framework that is specifically dedicated to (i) the extraction of rare itemsets and (ii) the generation of rare association rules. It is based on an intuitive yet formal definition of rare itemset and rare association rule. Our goal is to provide a theoretical foundation for rare pattern mining and rare association rule generation, with definitions of reduced representations and complexity results for mining tasks, as well as to develop an algorithmic tool suite (within the CORON project [25]) together with the guidelines for its use.

It is noteworthy that playing with minimal support is not the only way to approach the mining of atypical regularities. Thus, different statistical measures may be used to assess atypicality of patterns that are not bound to the number of occurrences. Moreover, the availability of an explicitly expressed body of expert knowledge or expectations/beliefs (e.g., as general rules) for a particular dataset or analysis problem enables a more focused pattern extraction where an unexpected or exceptional pattern is assessed with respect to a generally admitted one (a relevant discussion thereof may be found in [27]).

Rare itemsets, similarly to frequent ones, could be easily turned into rules, i.e. by splitting them into premise and conclusion subsets. The resulting rules are necessarily rare but their confidence would vary. Only rules of high confidence can be reasonably considered as regularities.

The extraction of rare itemsets and rules presents significant challenges for data mining algorithms [26]. In particular, algorithms designed for frequent itemset mining are inadequate for extracting rare association rules. Therefore, as it was argued in [25], new specific algorithms have to be designed. The problem with conventional frequent itemset mining approaches is that they have a (physical) limit on how low the minimum support can be set. We call this absolute limit the *barrier*: the barrier is the absolute minimum support value that is still manageable for a given frequent itemset mining algorithm in a given computing environment. The exact position (value) of the barrier depends on several variables, such as: (1) the database (size, density, highly- or weakly-correlated, etc.); (2) the platform (characteristics of the machine that is used for the calculation (CPU, RAM)); (3) the software (efficient / less efficient implementation), etc. Conventional search techniques are *always* dependent on a physical limit that cannot be crossed: it is almost certain that the minimum support cannot be lowered to 1.[1] The questions that arise are: how can the barrier be crossed; what is on the other side of the barrier; what kind of information is hidden; and mainly, how to extract interesting association rules from the negative side of the barrier.

1.3 Contribution

In order to generate rare association rules, first rare itemsets have to be extracted. In [18] it is stated that the negative border of frequent itemsets can be found with levelwise algorithms. In the next section, first we propose a straightforward modification of the *Apriori* algorithm for this task called *Apriori-Rare*. During the levelwise search, *Apriori* computes the support of *minimal rare itemsets* (mRIs), i.e. rare itemsets such that all proper subsets are frequent. Instead of pruning the mRIs, *Apriori-Rare* retains them. After *Apriori-Rare* we introduce an optimized method called *MRG-Exp* that limits the exploration to frequent generators only. Generators are itemsets that have no proper subsets with the same support. Experimental results reveal that *MRG-Exp* is more efficient on dense, highly correlated datasets. In addition, we show that the output of the two algorithms are *identical*.

In the second part of the paper, we focus on the search for valid rare association rules, i.e. rules with low support and high confidence. Once all rare itemsets are available, in theory it is possible to generate all valid rare association rules. However,

[1] When the absolute value of minimum support is 1, then all existing itemsets are frequent.

this method has two drawbacks. First, the restoration of all rare itemsets is a very memory-expensive operation due to the huge number of rare itemsets. Second, having restored all rare itemsets, the number of generated rules would be even more. Thus, the same problem as in the case of frequent valid association rules has to be faced: dealing with a huge number of rules of which many are redundant and not interesting at all.

Frequent itemsets have several condensed representations, e.g. closed itemsets [21], generators representation [13], free-sets [1], non-derivable itemsets [5], etc. However, from the application point of view, the most useful representations are closed itemsets and generators. Among frequent association rules, bases are special rule subsets from which all other frequent association rules can be restored with a proper inference mechanism. The set of minimal non-redundant association rules ($\mathcal{MNR}$) is particularly interesting, because it is a lossless, sound, and informative representation of all valid (frequent) association rules [14]. Moreover, these frequent rules allow one to deduce a maximum of information with minimal hypotheses. Accordingly, the same sort of subset has been searched for rare rules, namely the set of minimal rare itemset rules, presented hereafter.

The present work is motivated by the long-standing open question of devising an efficient algorithm for finding rules that have a high confidence together with a low support. This work shows a number of characteristics that are of importance. First, valid rare association rules can be extracted efficiently. Second, an interesting subset of rare association rules can be directly computed, similar to the set of (frequent) $\mathcal{MNR}$ rules in the case of frequent rules. Third, the method is rather easy to implement.

The paper is organized as follows. The basic concepts and definitions for frequent and rare itemsets together with the computationally motivated results are presented in Section 2. Our two methods for computing the minimal rare itemsets are included in the same section. Then, Section 3 details the generation of informative rare association rules from rare itemsets. A detailed experimental study of the algorithms is provided in Section 4. Finally, Section 5 concludes the paper.

2 Frequent and Rare Itemsets

Consider the following 5×5 sample dataset: $\boldsymbol{\mathcal{D}} = \{(1,\ ABDE),\ (2,\ AC),\ (3,\ ABCE),\ (4,\ BCE),\ (5,\ ABCE)\}$. Throughout the paper, we will refer to this example as **"dataset $\boldsymbol{\mathcal{D}}$"**.

2.1 Basic concepts

We consider a set of *objects* or *transactions* $\mathcal{O} = \{o_1, o_2, \ldots, o_m\}$, a set of *attributes* or *items* $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, and a relation $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$. A set of items is called an *itemset*. Each transaction has a unique identifier (*tid*), and a set of transactions is called a *tidset*. The tidset of all transactions sharing a given itemset X is its *image*, denoted $t(X)$. For instance, the image of $\{A, B\}$ in $\boldsymbol{\mathcal{D}}$ is $\{1, 3, 5\}$, i.e., $t(AB) = 135$ in our separator-free set notation. The *length* of an itemset X is $|X|$, whereas an itemset of length i is called an i-itemset. The (absolute) *support* of an itemset X, denoted by $supp(X)$, is the size of its image, i.e. $supp(X) = |t(X)|$.

Support is a prime measure of interest for itemsets: one is typically – but not exclusively – interested in regularities in the data that manifest in recurring patterns. Thus, intuitively, the itemsets of higher support are more attractive. Formally, the frequent itemset mining assumes a search space for interesting patterns that correspond to the Boolean lattice $\mathcal{B}(2^{\mathcal{A}})$ of all possible itemsets (see Figure 1). The lattice is separated into two segments or zones through a user-provided "minimum support" threshold, denoted by *min_supp*. Thus, given an itemset X, if $supp(X) \geqslant min\_supp$, then it is called *frequent*. Dually, if a maximal support threshold max_supp is provided then an itemset P such that $supp(P) \leqslant max\_supp$ is called *rare* (or *infrequent*).

Frequent itemsets (FIs) and rare itemsets belong to two mutually complementary subsets of the powerset $2^{\mathcal{A}}$ that further represent contiguous zones of the lattice $\mathcal{B}(2^{\mathcal{A}})$. In the technical language of lattice theory [6], these zones represent an *order ideal* (or *downset*) and an *order filter* (or *upset*), respectively, which means that a subset of a frequent itemset is necessarily frequent and, dually, a superset of a rare itemset is necessarily rare. In the lattice in Figure 1, the two zones corresponding to a support threshold of 3 are separated by a solid line. For example, the itemsets {A}, {AB}, or {BE} are frequent whereas {D}, {BD}, or {ACD} are rare.

The rare itemset family and the corresponding lattice zone is the target structure of our study. It may be further split into two parts, the itemsets of support zero, hereafter called *zero itemsets*[2] (X with $supp(X) = 0$), on the one hand, and all other rare itemsets, on the other hand. For instance, {BCD} is a zero itemset whereas {D} is a non-zero rare itemset.

It is noteworthy that the overall split of the lattice into three "stripes" depends for its exact shape on the chosen value for *min_supp*. Furthermore, it can be generalized to n stripes by providing an ordered sequence of $n-1$ values. Typically, we have assumed above that all itemsets can either be rare or frequent, but this needs not to always be the case. Thus, one can have two separate threshold values, one for each family, thus leaving a possibly void intermediate zone of neither-frequent-nor-rare itemsets.

Whatever the exact number of thresholds and zones, each zone is delimited by two subsets, the maximal elements and the minimal ones, respectively. For instance, the minimal frequent itemset is the empty set (whose support is $|\mathcal{D}|$) whereas the family of maximal frequent itemsets depends on *min_supp*. Similarly, the unique maximal rare itemset is $\mathcal{I}$ which is usually, but not invariably, a zero itemset.

The above intuitive ideas are formalized in the notion of a border introduced by Mannila and Toivonen in [18]. According to their definition, the maximal frequent itemsets constitute the *positive border* of the frequent zone whereas the minimal rare itemsets form the *negative border* of the same zone. Obviously, the same holds for the border between non-zero and zero itemsets as well.

Equivalence Class. An equivalence relation is induced by t on the power-set of items $\wp(\mathcal{A})$: equivalent itemsets share the same image ($X \cong Z$ iff $t(X) = t(Z)$) [4]. Consider the equivalence class of X, denoted $[X]$, and its extremal elements w.r.t. set inclusion. $[X]$ has a unique maximum (a *closed* itemset), and a set of minima (*generator* itemsets). A *singleton* equivalence class has only one element. The following definition exploits the monotony of support upon set inclusion in $\wp(\mathcal{A})$:

[2] Not to be confused with the empty set.

Definition 2.1. An itemset X is *closed* if it has no proper superset with the same support. An itemset Z is *generator* if it has no proper subset with the same support.

A *closure* operator underlies the set of closed itemsets; it assigns to X the maximum of $[X]$ (denoted by $\gamma(X)$). Naturally, $X = \gamma(X)$ for closed X. Generators, a.k.a. *key-sets* in database theory, represent a special case of free-sets [1]. The following property, which is widely known in the domain, basically states that the generator family is a downset within the Boolean lattice $\langle \wp(\mathcal{A}), \subseteq \rangle$:

Property 2.1. Given $X \subseteq \mathcal{A}$, if X is a generator, then $\forall Y \subseteq X$, Y is a generator, whereas if X is not a generator, $\forall Z \supseteq X$, Z is not a generator.

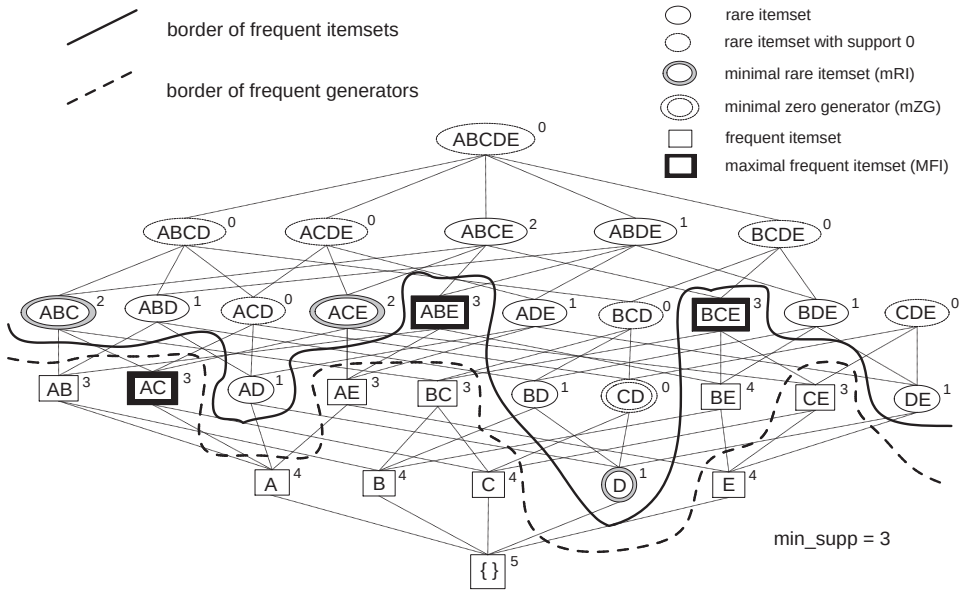

Figure 1. The powerset lattice of dataset $\mathcal{D}$.

2.2 Computationally motivated results

In order to ground an effective and efficient computation procedure for a particular zone, e.g., the frequent itemset family, one must provide a characterization of its members. Moreover, if the computation is done levelwise, i.e., by visiting iteratively lattice levels that are made of itemsets of a fixed size, one may also need a characterization of the zone border(s). Indeed, if the zone comprises none of the lattice extremal nodes, i.e., $\emptyset$ and $\mathcal{A}$, as is the case of the rare itemset zone, one needs to first pinpoint the starting points of the zone exploration. These starting points are typically the extremal elements, either maximal or minimal, i.e., the positive borders. Furthermore, the computation would typically need to traverse a neighbor zone, hence the negative border of the target zone must also be computed.

We consider here a computation of the rare itemsets that approaches them starting from the lattice bottom, i.e., from the frequent zone. Hence we need a characterization of what is widely known as the positive and the negative border of the frequent

itemsets, and corresponds for us to the negative *lower* border and the positive *lower* border of the rare itemsets, respectively. Moreover, should one need more than simply the rare itemsets on the border, the adverse *upper* border must be characterized as well.

First, the negative lower border of rare itemsets is a structure known from the literature. The characterization of its members, the *maximal frequent itemsets*, is straightforward:

Definition 2.2. An itemset is a *maximal frequent itemset* (MFI) if it is frequent but all its proper supersets are rare.

Second, the positive lower border of rare itemsets, i.e. the set of minimal rare itemsets is defined dually:

Definition 2.3. An itemset is a *minimal rare itemset* (mRI) if it is rare but all its proper subsets are frequent.

There are at least two possibilities for reaching the mRI family from the lattice bottom node that we discuss in the next subsections. On the one hand, as we indicated above, a levelwise search listing all frequent itemsets up to the MFIs represents a straightforward solution. Indeed, the levelwise search yields as a by-product all mRIs [18]. On the other hand, the computation of MFIs has been tackled by dedicated methods, hence an alternative solution will be to extract these itemsets directly and then use them as starting point in the computation of the mRIs, e.g., using the algorithm in [3]. The latter task is known to be computationally hard as it amounts to computing the minimal transversals of a hypergraph [2].

Hence we prefer a different optimization strategy that still yields mRIs while traversing only a subset of the frequent zone of the Boolean lattice. It exploits the minimal generator status of the mRIs. In Figure 1, the downset of frequent generators is delimited by a dashed line. For instance, knowing that {BC} is a frequent generator, {B} and {C} are necessarily frequent generators too. By Property 2.1, frequent generators (FGs) can be traversed in a levelwise manner while yielding their negative border as a by-product. Now, it is easy to see that all mRIs are part of the negative border of frequent generators. To that end, it is enough to observe that mRIs are in fact generators:

Proposition 2.1. All minimal rare itemsets are generators.

Thus, while there might well be other elements in the negative border that are not generators, e.g., frequent itemsets other than generators, all mRIs will necessarily lay on this border. More specifically, all the rare itemsets on that border will necessarily be minimal for their zone.

It remains now to provide an efficient criterion for recognizing frequent generators. The following property is a reduction of the initial definition to the immediate predecessors of a generator in the lattice (see [24]):

Proposition 2.2. An itemset X is a generator iff $supp(X) \neq min_{i \in X}(supp(X \setminus \{i\}))$.

The property says that in order to decide whether a candidate set X is a generator, one needs to compare its support to the support of its immediate predecessors in the lattice, i.e., the subsets of size $|X| - 1$. Obviously, generators do not admit predecessors of the same support.

The equivalence of the above results can be established for the upper border of the rare non-zero zone of the lattice. Thus, minimal zero generators can be defined as:

Definition 2.4. A *minimal zero generator* (mZG) is a zero itemset whose proper subsets are all non-zero itemsets.

For instance, in Figure 1 there is only one mZG element, {CD}. Finally, it is noteworthy that both sides of the border between frequent and rare itemsets play dual role in their respective zones. Indeed, beside being extremal elements, i.e., maximal and minimal, respectively, they constitute reduced representations for these zones as well. For instance, to extract the entire family of frequent itemsets from the MFIs, one only needs to generate all possible subsets thereof. Conversely, if all rare itemsets, i.e., zero and non-zero ones, are necessary, a dual technique will work that amounts to generating all supersets of mRIs [22]. Should zero itemsets be unnecessary, then minimal zero generators would work as stop criterion: only supersets of mRIs that do not include a minimal zero generator will be kept. Provided the support of these sets is required, it can be easily computed along a single pass through the database.

The next two subsections present the two methods for mRI computation.

2.3 Finding mRIs with a naïve approach

As pointed out by Mannila and Toivonen in [18], the easiest way to reach the negative border of the frequent itemset zone, i.e., the mRIs, is to use a levelwise algorithm such as Indeed, albeit a frequent itemset miner, *Apriori* yields the mRIs as a by-product. The mRIs are milestones in the exploration as they indicate that the border of the frequent zone has been crossed.

The overall principle of *Apriori* is rather intuitive: frequent itemsets are generated levelwise, at each iteration i targeting the itemset of length i, i.e., the i^{th} level above the lattice bottom node. The algorithm generates a set of candidates that are further matched against the database to evaluate their support in one database pass per iteration. To avoid redundant checks, two techniques are used: **(i)** candidates at level $i+1$ are generated by joining frequent i-itemsets that share $i-1$ of their items, thus increasing the chance of the result being frequent, and **(ii)** candidates are pruned *a priori*, i.e., before support computing, by eliminating those having a rare subset (of size $i-1$). In doing that, there is no need to explicitly represent rare itemsets: rather, all $i-1$ subsets of a candidate are generated dynamically and their presence in the frequent itemset storage structure is tested (absence means the subset, hence the candidate too, is rare).

Apriori-Rare is a slightly modified version of *Apriori* that stores the mRIs. Thus, whenever an i candidate survives the frequent $i-1$ subset test, but proves to be rare, it is kept as an mRI. For example, following the execution of *Apriori* on dataset $\mathcal{D}$, we get the following result. In C_1 (the set of 1-long candidates), there are 5 itemsets ({A}, {B}, {C}, {D}, and {E}) of which {D} is rare. In C_2 all itemsets are frequent ({AB}, {AC}, {AE}, {BC}, {BE}, and {CE}). In C_3 ({ABC}, {ABE}, {ACE}, and {BCE}) there are two rare itemsets namely {ABC} and {ACE}. Saving the three rare itemsets, one can obtain the following minimal rare itemsets at the end: {D}, {ABC}, and {ACE}.

Algorithm MRG-Exp:

Description: finding minimal rare generators efficiently
Input: dataset plus min_supp
Output: FGs plus mRGs

```
1)  CG_1 ← {1-itemsets};
2)  SupportCount(CG_1);   //requires one database pass
3)  loop over the rows of CG_1 (c) {
4)      c.pred_supp ← ∅.supp;   //i.e., c.pred_supp ← |O|;
5)      if (c.pred_supp = c.supp) c.key ← false;
6)      else                      c.key ← true;
7)  }
8)  RG_1 ← { r ∈ CG_1 | (r.key=true) ∧ (r.supp < min_supp) };
9)  FG_1 ← { f ∈ CG_1 | (f.key=true) ∧ (f.supp ⩾ min_supp) };
10) for (i ← 1; true; i ← i + 1)
11) {
12)     CG_{i+1} ← GenCandidates(FG_i);
13)     if (CG_{i+1} = ∅) break;   //i.e., break out from the "for" loop
14)     SupportCount(CG_{i+1});   //requires one database pass
15)     loop over the rows of CG_{i+1} (c)
16)     {
17)         if (c.pred_supp != c.supp) {   //i.e., if c is a generator
18)             if (c.supp < min_supp) RG_{i+1} ← RG_{i+1} ∪ {c};
19)             else                   FG_{i+1} ← FG_{i+1} ∪ {c};
20)         }
21)     }
22) }
23) G_F ← ⋃_i FG_i;      //frequent generators
24) G_MR ← ⋃_i RG_i;   //minimal rare generators
```

2.4 Finding mRIs in an efficient way

Following Proposition 2.1, we may avoid exploring all frequent itemsets: instead, it is sufficient to look after frequent generators only. In this case, mRIs, which are rare generators as well, can be filtered among the negative border of the frequent generators.

For finding minimal rare generators, we focus exclusively on frequent generators and their downset in the lattice (see Algorithm *MRG-Exp*). Thus, frequent i-long generators are joined to create $(i+1)$-long candidates. These undergo a series of tests. On the one hand, the generator status is established following Proposition 2.2 with the additional condition that all subsets of the candidate must be frequent generators. Thus, non-generator frequent itemsets and non-minimal rare itemsets are discarded. Next, frequency test against the database is used to separate frequent from (minimal) rare generators.

CG_1	pred_supp	key	supp
{A}	5	yes	4
{B}	5	yes	4
{C}	5	yes	4
{D}	5	yes	1
{E}	5	yes	4

RG_1	supp
{D}	1

FG_1	supp
{A}	4
{B}	4
{C}	4
{E}	4

CG_2	pred_supp	key	supp
{AB}	4	yes	3
{AC}	4	yes	3
{AE}	4	yes	3
{BC}	4	yes	3
{BE}	4	—	4
{CE}	4	yes	3

RG_2	supp
∅	

FG_2	supp
{AB}	3
{AC}	3
{AE}	3
{BC}	3
{CE}	3

CG_3	pred_supp	key	supp
{ABC}	3	yes	2
{ABE}	3	—	3
{ACE}	3	yes	2

RG_3	supp
{ABC}	2
{ACE}	2

FG_3	supp
∅	

CG_4	pred_supp	key	supp
∅			

Figure 2. Execution of the MRG-Exp algorithm.

The above reasoning is partly embedded into the `GenCandidates` function which has three-fold effect. First, it produces the $(i+1)$-long candidate generators, using the i-long frequent generators in the FG_i table. Second, all candidates having an i-long subset which is not in FG_i are deleted. In this way, *non-minimal* rare itemsets are pruned, and only potential generators are kept. Third, the function determines the pred_supp values of the candidates, i.e., the minimum of the supports of all i-long subsets.

Later in the process, the pred_supp is compared to the actual support of a candidate. If both values are different then the candidate is a true generator. Moreover, depending on its support, it is either a frequent generator or a minimal rare one, i.e., an mRI.

The execution of *MRG-Exp* on dataset $\mathcal{D}$ with $min\_supp = 3$ is illustrated in Figure 2. The algorithm first performs one database scan to count the supports of 1-long itemsets. The pred_supp column indicates the minimum of the supports of all $(i-1)$-long frequent subsets. Itemsets of length 1 only have one frequent subset, the empty set. By definition, the empty set is included in every object of the database, thus its support is 100%. Comparing the support and pred_supp values, it turns out that all 1-itemsets are generators. Testing the support values, itemset {D} is copied to RG_1, while the other generators are copied to FG_1. In CG_2 there is one itemset that has the same support as one of its subsets, thus {BE} is not a (key) generator. In the fourth iteration no new candidate is found and the algorithm breaks out from the

main loop. When the algorithm stops, all minimal rare generators are found ({D}, {ABC}, and {ACE}).

2.5 Complexity of the mRG computation

The theoretical complexity of the above algorithm is bound to the complexity of the levelwise algorithms for frequent itemset mining. Thus, due to the potentially exponential size of the output, there is no point in establishing a conventional estimation thereof in terms of the O-based notation. Indeed, in the worst case there will be exponentially many FGs, hence any comparable algorithm will have an exponential complexity function. Therefore, a more reasonable measure for efficiency would be provided by the computational cost per single generator, the amount of work to compute a single member of the entire FG/mRG family. Following [7], it is easy to see that this quantity is bounded by a polynomial function of the following factors: (1) the maximal size of a mRG/FG, (2) the size of the transaction database, and (3) the number of items. As of the complexity class of the algorithm, it is necessarily in the total polynomial class, following the classification of [9] for algorithms that list all the solutions of a decision problem. The stronger notion of polynomial delay, meaning that the delay between any two outputs of the algorithm (mRG) is polynomial in the size of the input, is also satisfied. This is an important quality as such algorithms take time linear in the combined size of their input and output.

3 Rare Association Rules

3.1 Basic concepts

An association rule is an expression of the form $P_1 \rightarrow P_2$, where P_1 and P_2 are arbitrary itemsets ($P_1, P_2 \subseteq \mathcal{A}$), $P_1 \cap P_2 = \emptyset$ and $P_2 \neq \emptyset$. The left side, P_1 is called *antecedent*, the right side, P_2 is called *consequent*. The support of an association rule r: $P_1 \rightarrow P_2$ is defined as: $supp(r) = supp(P_1 \cup P_2)$. The *confidence* of an association rule r: $P_1 \rightarrow P_2$ is defined as the conditional probability that an object includes P_2, given that it includes P_1: $conf(r) = supp(P_1 \cup P_2)/supp(P_1)$. An association rule r is called *confident*, if its confidence is not less than a given *minimum confidence* (denoted by min_conf), i.e. $conf(r) \geqslant min\_conf$. An association rule r with $conf(r) = 1.0$ (i.e. 100%) is an *exact* association rule, otherwise it is an *approximate* association rule.

An association rule r is called *frequent* if its support is not less than a given *minimum support* (denoted by min_supp), i.e. $supp(r) \geqslant min\_supp$. A frequent association rule is *valid* if it is confident, i.e. $supp(r) \geqslant min\_supp$ and $conf(r) \geqslant min\_conf$. *Minimal non-redundant association rules* ($\mathcal{MNR}$) have the following form: $P \rightarrow Q \setminus P$, where $P \subset Q$ and P is a frequent *generator* and Q is a frequent *closed* itemset.

An association rule is called *rare* if its support is not more than a given *maximum support*. Since we use a single border, it means that a rule is rare if its support is less than a given *minimum support*. A rare association rule r is *valid* if r is confident, i.e. $supp(r) < min\_supp$ and $conf(r) \geqslant min\_conf$. In the rest of the paper, by "rare association rules" we mean *valid* rare association rules.

3.2 Breaking the barrier

Recall that our goal is to break the *barrier*, i.e. to be able to extract rare itemsets and rare association rules that cannot be extracted with the direct approach used by conventional frequent itemset mining algorithms like *Apriori*. With the *BtB* (Breaking the Barrier) algorithm we can extract highly confident rare association rules below the barrier. The algorithm consists of the following three main steps.

First, for computing the set of minimal rare itemsets, the key algorithm is *MRG-Exp*. *MRG-Exp* finds frequent generators, but as a "side effect" it also explores the so-called minimal rare generators (mRGs). *MRG-Exp* retains these itemsets instead of pruning them. In Section 2.2 we show that the set of minimal rare itemsets is identical to the set of minimal rare generators (see Proposition 2.1).

Second, find the closures of the previously found minimal rare generators so as to obtain their equivalence classes.

Third, from the explored rare equivalence classes it is possible to generate rare association rules in a way very similar to that of finding (frequent) minimal non-redundant association rules. We call these rare rules "mRG rules" because their antecedents are minimal rare generators.

3.3 mRG rules

Two kinds of mRG rules can be distinguished, namely exact and approximate rules. In this paper we concentrate on exact mRG rules that can be characterized as:

$$r\colon\ P_1 \Rightarrow P_2 \setminus P_1\ \text{, where} \qquad \begin{array}{l} P_1 \subset P_2 \\ P_1 \text{ is an mRG} \\ P_1 \cup (P_2 \setminus P_1) = P_2 \text{ is a rare closed itemset} \\ conf(r) = 1.0 \end{array}$$

From the form of exact mRG rules it follows that these rules are *rare* association rules, where the antecedent (P_1) is rare and the consequent ($P_2 \setminus P_1$) is rare *or* frequent. P_1 and P_2 are in the same equivalence class.

Since a generator is a minimal subset of its closure with the same support, these rules allow us to deduce maximum information with minimal hypothesis, just as the $\mathcal{MNR}$ rules. Using Kryszkiewicz's cover operator [12], one can restore further *exact* rare association rules from the set of exact mRG rules.

Example. Figure 3 shows all the equivalence classes of dataset $\mathcal{D}$. Support values are depicted above to the right of equivalence classes. Itemsets with the same support are grouped together in the same level. Levels are separated by borders that are defined by different *min_supp* values. Next to each *min_supp* value, the corresponding minimal rare itemsets are also shown. For instance, if $min\_supp = 4$ then there exist 5 frequent itemsets (A, C, B, E, BE) and 6 minimal rare itemsets (D, AB, AC, AE, BC, CE).

Suppose that the barrier is at $min\_supp = 4$. In this case, using *Apriori*, the less frequent association rules have support 4. With *Apriori-Rare* or *MRG-Exp*, the following mRIs are found: D, AB, AC, AE, BC and CE. Calculating their closures, four rare equivalence classes are explored, as shown in Figure 4 (left). Note that *not all* rare equivalence classes are found. For instance, the class whose maximal element

is $ABCE$ is not found because its generators are *not* mRIs, i.e. it is not true for ABC and ACE that all their proper subsets are frequent itemsets.

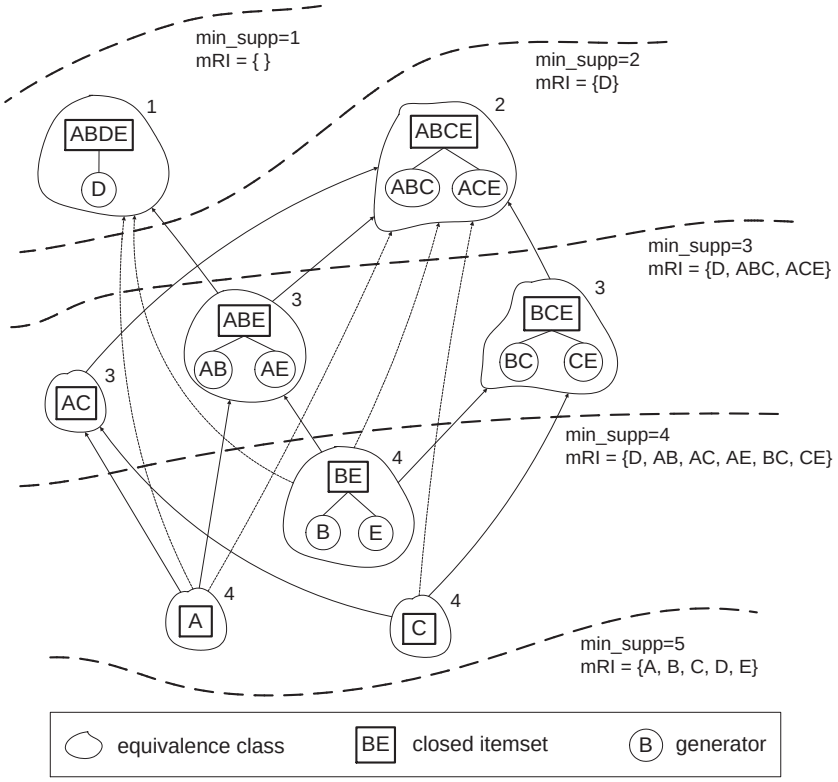

Figure 3. Rare equivalence classes found by *BtB* in dataset $\mathcal{D}$ at different *min_supp* values.

Generating exact mRG rules. Once rare equivalence classes are found, the rule generation method is basically the same as in the case of $\mathcal{MNR}$ rules. Exact mRG rules are extracted within the same equivalence class. Such rules can only be extracted from non-singleton classes. Figure 4 (center) shows which exact mRG rules can be extracted from the found rare equivalence classes (Figure 4, left).

Generating approximate mRG rules. Approximate mRG rules are extracted from classes whose maximal elements are comparable with respect to set inclusion. Let P_1 be an mRG, $\gamma(P_1)$ the closure of P_1, and $[P_1]$ the equivalence class of P_1. If a proper superset P_2 of $\gamma(P_1)$ is picked among the maximal elements of the found rare equivalence classes different from $[P_1]$, then $P_1 \rightarrow P_2 \backslash P_1$ is an approximate mRG rule. Figure 4 (right) shows the approximate mRG rules that can be extracted from the found rare equivalence classes (Figure 4, left).

4 Experimental Results

In this section we present the results of a series of tests. First, we compare the performances of *Apriori-Rare* and *MRG-Exp*. Then, we provide results that we obtained on a real-life biomedical dataset. Finally, we demonstrate that our approach

closure	supp.	generators
$ABDE$	1	D
AC	3	AC
ABE	3	AB, AE
BCE	3	BC, CE

rule	supp.	conf.
$D \Rightarrow ABE$	1	1.0
$AB \Rightarrow E$	3	1.0
$AE \Rightarrow B$	3	1.0
$BC \Rightarrow E$	3	1.0
$CE \Rightarrow B$	3	1.0

rule	supp.	conf.
$AB \rightarrow DE$	1	1/3
$AE \rightarrow BD$	1	1/3

Figure 4. **Left:** rare equivalence classes found by *BtB* in $\mathcal{D}$ with $min\_supp = 4$. **Center:** exact mRG rules in $\mathcal{D}$ with $min\_supp = 4$. **Right:** approximate mRG rules in $\mathcal{D}$ with $min\_supp = 4$.

is computationally efficient for extracting rare itemsets and rare association rules. Thus, a series of computational times resulting from the application of our algorithms to well-known datasets is presented.

The algorithms were implemented in Java in the CORON platform [25].[3] The experiments were carried out on an Intel Pentium IV 2.4 GHz machine running under Debian GNU/Linux operating system with 512 MB RAM. All times reported are real, wall clock times as obtained from the Unix *time* command between input and output.

For the experiments we have used the following datasets: T20I6D100K, C20D10K, C73D10K, and MUSHROOMS. Database characteristics are shown in Table 1. The T20I6D100K[4] is a sparse dataset, constructed according to the properties of market basket data that are typical weakly correlated data. The C20D10K and C73D10K are census datasets from the PUMS sample file, while the MUSHROOMS[5] describes mushrooms characteristics. The last three are dense, highly correlated datasets.

Table 1. Database characteristics

database name	# records	# non-empty attributes	# attributes (in average)	largest attribute
T20I6D100K	100,000	893	20	1,000
C20D10K	10,000	192	20	385
C73D10K	10,000	1,592	73	2,177
MUSHROOMS	8,416	119	23	128

4.1 Apriori-Rare vs. MRG-Exp

In our experiments we compared *Apriori-Rare* and *MRG-Exp*. The execution times of the two algorithms are illustrated in Table 2. The table also shows the number of frequent itemsets, the number of frequent generators, the proportion of the number of FGs to the number of FIs, and the number of minimal rare itemsets.

The T20I6D100K synthetic dataset mimics market basket data that are typical sparse, weakly correlated data. In this dataset, the number of FIs is small and nearly all FIs are generators. Thus, *MRG-Exp* works exactly like *Apriori-Rare*, i.e. it has to explore almost the same search space. The reason why *MRG-Exp* is a bit slower is that *MRG-Exp* determines in addition the pred_supp value of each candidate generator.

In datasets C20D10K, C73D10K, and MUSHROOMS, the number of FGs is much less than the total number of FIs. Hence, *MRG-Exp* can take advantage of explor-

[3] http://coron.loria.fr
[4] http://www.almaden.ibm.com/software/quest/Resources/
[5] http://kdd.ics.uci.edu/

ing a much less search space than *Apriori-Rare*. Thus, *MRG-Exp* performs much better on dense, highly correlated data. For example, on the dataset MUSHROOMS at $min\_supp = 10\%$, *Apriori-Rare* needs to extract 600,817 FIs, while *MRG-Exp* extracts 7,585 FGs only. This means that *MRG-Exp* reduces the search space of *Apriori-Rare* to 1.26%!

Table 2. Response times of Apriori-Rare and MRG-Exp

min_supp	execution time (sec.)		# FIs	# FGs	$\frac{\#FGs}{\#FIs}$	# mRIs
	Apriori-Rare	MRG-Exp				
T20I6D100K						
10%	11.47	15.91	7	7	100.00%	907
0.75%	146.61	156.65	4,710	4,710	100.00%	211,578
0.5%	238.27	262.32	26,836	26,305	98.02%	268,915
0.25%	586.21	622.30	155,163	149,447	96.32%	537,765
C20D10K						
30%	125.97	26.55	5,319	967	18.18%	230
20%	326.87	50.31	20,239	2,671	13.20%	400
10%	842.85	104.25	89,883	9,331	10.38%	901
5%	1,785.08	162.07	352,611	23,051	6.54%	2,002
2%	4,074.33	228.44	1,741,883	57,659	3.31%	7,735
C73D10K						
95%	216.04	37.04	1,007	121	12.02%	1,622
90%	2,567.42	253.08	13,463	1,368	10.16%	1,701
85%	9,364.20	607.85	46,575	3,513	7.54%	1,652
MUSHROOMS						
40%	13.73	6.00	505	153	30.30%	254
30%	46.10	12.64	2,587	544	21.03%	409
15%	869.27	40.68	99,079	3,084	3.11%	1,846
10%	3,097.16	69.23	600,817	7,585	1.26%	3,077

4.2 The Stanislas cohort

A cohort study consists of examining a given population during a period of time and of recording different data concerning this population. Data from a cohort show a high rate of complexity: they vary in time, involve a large number of individuals and parameters, show many different types, e.g. quantitative, qualitative, textual, binary, etc., and they may be noisy or incomplete. whose main objective is to investigate the impact of genetic and environmental factors on variability of cardiovascular risk factors [17]. The cohort consists of 1006 presumably healthy families (4295 individuals) satisfying some criteria: French origin, two parents, at least two biological children aged of 4 or more, with members free from serious and/or chronic illnesses. The collected data are of four types: (1) Clinical data (e.g. size, weight, blood pressure); (2) Environmental data (life habits, physical activity, drug intake); (3) Biological data (glucose, cholesterol, blood count); (4) Genetic data (genetic polymorphisms).

The experts involved in the study of the STANISLAS cohort are specialists of the cardiovascular domain and they are interested in finding associations relating one or more genetic features (polymorphisms) to biological cardiovascular risk factors. The objective of the present experiment is to discover rare association rules linking biolog-

ical risk factors and genetic polymorphisms. As a genetic polymorphism is defined as a variation in the DNA sequence occurring in at least one percent of the population, it is easily understandable that the frequency of the different genetic variants is relatively low in the STANISLAS cohort, given that it is based on a healthy population. Therefore, this fully justifies an analysis based on rare association rules [25].

Here is an example of the extraction of a new biological hypothesis derived from the study of the STANISLAS cohort. The objective of the experiment is to characterize the genetic profile of individuals presenting "metabolic syndrome" (depending on criteria such as waist circumference, triglyceride levels, HDL cholesterol concentration, blood pressure, and fasting glucose value). A horizontal projection allowed us to retain nine individuals with metabolic syndrome. Then, a vertical projection was applied on a set of chosen attributes. Rare association rules were computed and the set of extracted rules was mined for selecting rules with the attribute *metabolic syndrome* in the left or in the right hand side. In this way, an interesting extracted rule has been discovered: `MS` $\Rightarrow$ `APOB_71ThrIle` (support 9 and confidence 100%). This rule can be interpreted as "an individual presenting the metabolic syndrome is heterozygous for the APOB 71Thr/Ile polymorphism". This rule has been verified and validated using statistical tests, allowing us to conclude that the repartition of genotypes of the APOB71 polymorphism is significantly different when an individual presents metabolic syndrome or not, and suggests a new biological hypothesis: a subject possessing the rare allele for the APOB 71Thr/Ile polymorphism presents more frequently the metabolic syndrome. Other examples of rare rules can be found in [25].

4.3 Further experiments

We evaluated *BtB* on the four datasets mentioned before. Table 3 shows the different steps of finding exact mRG rules. The table contains the following columns: (1) Name of the dataset and minimum support values; (2) Number of frequent itemsets. It is only indicated to show the combinatorial explosion of FIs as *min_supp* is lowered; (3) Number of mRGs whose support exceeds 0. Since the total number of zero itemsets can be huge, we have decided to prune itemsets with support 0; (4) Number of non-singleton rare equivalence classes that are found by using non-zero mRGs; (5) Number of found exact (non-zero) mRG rules; (6) Total runtime of the *BtB* algorithm, including input/output.

During the experiments we used two limits: a space limit, which was determined by the main memory of our test machine, and a time limit that we fixed as 10,000 seconds. The value of the barrier is printed in bold in Table 3. For instance, in the database C73D10K using *Apriori* we were unable to extract any association rules with support lower than 65% because of hitting the time limit. However, changing to *BtB* at this *min_supp* value, we managed to extract 3,675 exact mRG rules whose supports are *below* 65%. This result shows that our method is capable to find rare rules where frequent itemset mining algorithms fail.

Table 3. Steps taken to find the exact mRG association rules

dataset and min_supp	# FIs	# mRGs (non-zero)	# rare eq. classes (non-zero, non-singleton)	**# mRG rules (exact)**	runtime of the BtB alg. (sec.)
$\mathcal{D}$, 80%	5	6	3	5	0.09
T20I6D100K, 10%	7	907	27	27	25.36
0.75%	4,710	211,561	4,049	4,053	312.63
0.5%	26,836	268,589	16,100	16,243	742.40
0.25%	155,163	534,088	43,458	45,991	2,808.54
C20D10K, 10%	89,883	837	778	837	102.09
1%	6,194,967	15,433	12,485	15,433	302.97
0.5%	15,602,883	33,266	25,165	33,266	401.41
0.25%	40,450,371	62,173	41,915	62,173	640.95
C73D10K, 95%	1,007	1,622	1,570	1,622	59.10
75%	235,271	1,939	1,794	1,939	2,183.70
70%	572,087	2,727	2,365	2,727	4,378.02
65%	1,544,691	3,675	2,953	3,675	9,923.94
MUSHROOMS, 50%	163	147	139	147	3.38
10%	600,817	2,916	2,324	2,916	74.60
5%	4,137,547	7,963	5,430	7,963	137.86
1%	92,894,869	37,034	16,799	37,034	321.78

5 Conclusion

Frequent association rule mining has been studied extensively in the past. The model used in all these studies, however, has always been the same, i.e. finding all rules that satisfy user-specified *min_supp* and *min_conf* constraints. However, in many cases, most rules with high support are obvious and/or well-known, and it is the rules of low support that provide interesting new insights.

In the first part of the paper, we presented an approach for rare itemset mining from a dataset. The traversal of the frequent zone in the space is addressed by two different algorithms, a naïve one, *Apriori-Rare*, which relies on *Apriori* and hence enumerates *all* frequent itemsets; and an optimized one, *MRG-Exp*, which limits the considerations to frequent generators *only*. Experimental results prove the interest of the optimized method on dense, highly correlated datasets.

In the second part of the paper, we presented a novel method to extract interesting rare association rules that remain *hidden* for conventional frequent itemset mining algorithms. To the best of our knowledge, this is the first method in the literature that can find strong but rare associations, i.e., local regularities in the data. These rules, called "mRG rules", have two merits. First, they are maximally informative in the sense that they have an antecedent which is a generator itemset whereas adding the consequent to it yields a closed itemset. Second, the number of these rules is minimal, i.e. the mRG rules constitute a compact representation of all highly confident associations that can be drawn from the minimal rare itemsets.

References

[1] Boulicaut JF, Bykowski A, Rigotti C. Free-Sets: A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. Data Mining and Knowledge Discovery, 2003, 7(1): 5–22.

[2] Berge C. Hypergraphs: Combinatorics of Finite Sets. North Holland, Amsterdam, 1989.

[3] Boros E, Gurvich V, Khachiyan L, et al. On maximal frequent and minimal infrequent sets in binary matrices. Annals of Mathematics and Artificial Intelligence, 2003, 39: 211-221.

[4] Bastide Y, Taouil R, Pasquier N, et al. Mining Frequent Patterns with Counting Inference. SIGKDD Explor. Newsl., 2000, 2(2): 66–75.

[5] Calders T, Goethals B. Mining All Non-derivable Frequent Itemsets. Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '02). London, UK, 2002. Springer-Verlag. 2002. 74–85.

[6] Davey BA, Priestley HA. Introduction to Lattices and Order. Cambridge University Press, 2nd edition, 2002.

[7] Gunopulos D, Khardon R, Mannila H, et al. Discovering all most specific sentences. ACM Trans. on Database Systems, 2003, 28(2): 140–174.

[8] Ganter B, Wille, R. Formal Concept Analysis: Mathematical Foundations. Springer, Berlin/Heidelberg, 1999.

[9] Johnson DS, Papadimitriou CH. On generating all maximal independent sets. Information Processing Letters, 1988, 27(3): 119–123.

[10] Koh YS, Rountree N. Finding Sporadic Rules Using Apriori-Inverse. Proc. of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '05). Hanoi, Vietnam, LNCS3518. Springer, May 2005. 97–106.

[11] Koh YS, Rountree N, O'Keefe R. Mining Interesting Imperfectly Sporadic Rules. Proc. of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '06), Singapore. LNCS3918. Springer, April 2006. 473–482.

[12] Kryszkiewicz M. Representative Association Rules. Proc. of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD '98). Melbourne, Australia, 1998. Springer-Verlag. 1998. 198–209.

[13] Kryszkiewicz M. Concise Representation of Frequent Patterns Based on Disjunction-Free Generators. Proc. of the 2001 IEEE International Conference on Data Mining (ICDM '01). Washington, DC, 2001. IEEE Computer Society. 2001. 305–312.

[14] Kryszkiewicz M. Concise Representations of Association Rules. Proc. of the ESF Exploratory Workshop on Pattern Detection and Discovery. 2002. 92–109.

[15] Liu B, Hsu W, Ma Y. Mining Association Rules with Multiple Minimum Supports. Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99). New York, NY, USA, 1990. ACM Press. 1990. 337–341.

[16] Liu H, Lu H, Feng L, et al. Efficient Search of Reliable Exceptions. Proc. of the 3rd Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining (PAKDD '99). London, UK, 1999. Springer-Verlag. 1999. 194–203.

[17] Mansour-Chemaly M, Haddy N, Siest G. Family studies: their role in the evaluation of genetic cardiovascular risk factors. Clin. Chem. Lab. Med., 2002. 40(11): 1085–1096.

[18] Mannila H, Toivonen H. Levelwise Search and Borders of Theories in Knowledge Discovery. Data Mining and Knowledge Discovery, 1997, 1(3): 241–258.

[19] Novak PK, Lavrac N, Webb GI. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research, 2009, 10: 377–403.

[20] Okubo Y, Haraguchi M. An Algorithm for Extracting Rare Concepts with Concise Intents. 8th Intl. Conf. on Formal Concept Analysis (ICFCA '10). LNCS5986, Agadir, Morocco, 2010. Springer. 2010. 145–160.

[21] Pasquier N, Bastide Y, Taouil R, et al. Efficient mining of association rules using closed itemset lattices. Inf. Syst., 1999, 24(1): 25–46.

[22] Szathmary L, Maumus S, Petronin P, et al. Vers l'extraction de motifs rares. In: Ritschard G,

Djeraba C, eds. Extraction et gestion des connaissances – EGC '06, Lille, France. RNTI-E-6, Cépaduès-Éditions Toulouse, 2006. 499–510.

[23] Szathmary L, Napoli A, Valtchev P. Towards Rare Itemset Mining. Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '07). Patras, Greece, Oct 2007. 305–312.

[24] Szathmary L, Valtchev P, Napoli A. Finding Minimal Rare Itemsets and Rare Association Rules. Proc. of the 4th Intl. Conf. on Knowledge Science, Engineering & Management (KSEM '10), vol. 6291 of LNAI, pages 16–27, Belfast, Northern Ireland, UK, 2010. Springer, Berlin.

[25] Szathmary L. Symbolic Data Mining Methods with the Coron Platform. PhD Thesis in Computer Science, Université Henri Poincaré – Nancy 1, France, Nov 2006.

[26] Weiss GM. Mining with rarity: a unifying framework. SIGKDD Explor. Newsl., 2004, 6(1): 7–19.

[27] Wang K, Jiang YL, Lakshmanan Laks VS. Mining unexpected rules by pushing user dynamics. Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03). New York, NY, USA, 2003. ACM. 2003. 246–255.

[28] Wu X, Zhang C, Zhang S. Efficient Mining of Both Positive and Negative Association Rules. ACM Trans. on Information Systems, 2004, 22(3): 381–405.

[29] Yun H, Ha D, Hwang B, et al. Mining association rules on significant rare data using relative support. Journal of Systems and Software, 2003, 67(3): 181–191.