Finding Minimal Rare Itemsets and Rare Association Rules

Laszlo Szathmary, Petko Valtchev, Amedeo Napoli

Finding Minimal Rare Itemsets and Rare Association Rules

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
Szathmary.L@gmail.com, valtchev.petko@uqam.ca
[2] LORIA UMR 7503, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France
napoli@loria.fr

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.

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 [1]. 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.

1.1 Motivating Examples

A first case study for atypical patterns and rules pertains to a French biomedical database, the Stanislas cohort [2]. The Stanislas cohort comprises the medical 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 event 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 Approaches and Recent Progress

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 [3].

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* [4,5]).

A less uniform support criterion is designed in [5] 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) [4] 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.

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* [6], and *MIISR* (Mining Interesting Imperfectly Sporadic Rules) [7] 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.

In our own approach, we concentrate on the dual part of the frequent subfamily, i.e., on all rare itemsets and not merely the perfectly rare ones. To that end, we devised a strategy that traverses the frequent zone of the itemset lattice (the order ideal of the frequent itemsets) at minimal cost, as described in [8]. The current paper is a follow-up dealing with rare rule generation out of the resulting set of rare itemsets (see next section).

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 [9]).

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 [3]. In particular, algorithms designed for frequent itemset mining are inadequate for extracting rare association rules. Therefore, 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 [10] it is stated that the negative border of frequent itemsets can be found with levelwise algorithms. A straightforward modification of the *Apriori* algorithm has been proposed in [8] for this task. 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, they are retained. In addition, it is shown that the mRIs form a generator set of rare itemsets, i.e. *all rare itemsets* can be restored from the set of mRIs [8].

In this 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, 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, generators representation, free-sets, non-derivable itemsets, 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 [11]. Moreover, these

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

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 are presented in Section 2. Then, Section 3 details the generation of informative rare association rules from rare itemsets. Finally, Section 4 concludes the paper.

2 Frequent and Rare Itemsets

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

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 $\mathcal{D}$ 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. $supp(X) = |t(X)|$. Moreover, X is *frequent*, if its support is not less than a given *minimum support* threshold min_supp, i.e. $supp(X) \geq min\_supp$. Dually, if a maximal support threshold max_supp is provided then an itemset P such that $supp(P) \leq max\_supp$ is called *rare* (or *infrequent*). If the support of an itemset is 0 then the itemset is a *zero itemset*[2], otherwise it is a *non-zero itemset*.

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 $t(X) = t(Z)$) [12]. 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 $\wp(\mathcal{A})$:

Definition 1. *An itemset X is* closed (generator) *if it has no proper superset (subset) with the same support.*

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

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 [13]. 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 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.

The separation of $\wp(\mathcal{A})$ into rare and frequent parts induces substructures that reflect the same extremum principle as generators/closures but in a global scope rather than within a single equivalence class.

Definition 2. ***(i)*** *A frequent itemset is a* maximal frequent itemset (MFI) *if all its proper supersets are not frequent.* ***(ii)*** *An itemset is a* minimal rare itemset (mRI) *if it is rare, and all its proper subsets are not rare.* ***(iii)*** *A* minimal rare generator (mRG) *is a rare generator such that and all its proper subsets are not rare.*

In [8] we showed that mRIs are in fact generators, i.e. the set of mRIs and the set of mRGs are equivalent:

Proposition 1. *All minimal rare itemsets are generators [8].*

In the general problem settings, an interval may exist between the thresholds min_supp and max_supp. Yet throughout the paper we shall assume that both values describe a unique separation of $\wp(\mathcal{A})$ into a frequent and a rare part (i.e. there will be no itemsets that are neither rare nor frequent). This basically means, in absolute terms, that $max\_supp = min\_supp - 1$.

The above equality amounts to the existence of cut across the powerset lattice separating the frequent part from the rare one. This cut, called hereafter the *border* as in [10], has a positive side, made of the frequent itemsets, and a negative side, made of the rare itemsets. Both sides of the border have intriguing mathematical properties (see [13,14]) whereas their computation has been reduced to well-known combinatorial generation problems [15].

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. $conf(r) \geq min\_conf$. 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. $supp(r) \geq min\_supp$. A frequent association rule is *valid* if it is confident, i.e. $supp(r) \geq min\_supp$ and $conf(r) \geq min\_conf$. *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 $conf(r) \geq min\_conf$. 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 *Apriori-Rare* [8]. *Apriori* finds frequent itemsets, but as a "side effect" it also explores the so-called minimal rare itemsets (mRIs). *Apriori-Rare* retains these itemsets instead of pruning them. In Section 2 we show that minimal rare itemsets are rare generators (see Proposition 1).

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

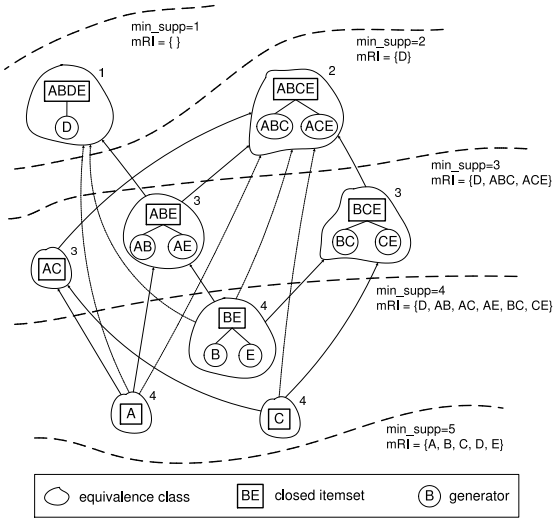
Third, from the explored rare equivalence classes it is possible to generate rare association rules in a way very similar to that of finding (frequent) minimal non-redundant association rules. We call these rare rules "mRG rules" because their antecedents are minimal rare generators.

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:

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

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.



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

Fig. 1. Left: rare equivalence classes found by *BtB* in dataset $\mathcal{D}$ at different min_supp values. **Top right:** rare equivalence classes found by *BtB* in $\mathcal{D}$ with $min\_supp = 4$. **Center right:** exact mRG rules in $\mathcal{D}$ with $min\_supp = 4$. **Bottom right:** approximate mRG rules in $\mathcal{D}$ with $min\_supp = 4$.

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 [16], one can restore further *exact* rare association rules from the set of exact mRG rules.

Example. Figure 1 (left) 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*, 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 1 (top right). 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.

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 1 (center right) shows which exact mRG rules can be extracted from the found rare equivalence classes (Figure 1, top right).

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 $P_1 \rightarrow P_2 \setminus P_1$ is an approximate mRG rule. Figure 1 (bottom right) shows the approximate mRG rules that can be extracted from the found rare equivalence classes (Figure 1, top right).

3.4 Experimental Results

In this section we present the results of a series of tests. First, we provide results that we obtained on a real-life biomedical dataset. Then, we demonstrate that our approach 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. All 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. Algorithms were implemented in the CORON platform [17].[3] All times reported are real, wall clock times; given in seconds.

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.

The STANISLAS cohort is a ten-year family study whose main objective is to investigate the impact of genetic and environmental factors on variability of cardiovascular risk factors [2]. 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

[3] http://coron.loria.fr

factors. The objective of the present experiment is to discover rare association rules linking biological 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 [17].

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 [17].

Further Experiments

We evaluated *BtB* on three more datasets. The T20I6D100K[4] is a sparse dataset, constructed according to the properties of market basket data that are typically sparse, weakly correlated data. The C73D10K is a census dataset from the PUMS sample file, while the MUSHROOMS[5] describes the characteristics of various species of mushrooms. The latter two are dense, highly correlated datasets. Table 1 shows the different steps of finding exact mRG rules. The table contains the following columns: (1) Name of the dataset; (2) Minimum support value; (3) Number of frequent itemsets. It is only indicated to show the combinatorial explosion of FIs as *min_supp* is lowered; (4) 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; (5) Number of non-singleton rare equivalence classes that are found by using non-zero mRGs; (6) Number of found exact (non-zero) mRG rules; (7) Total runtime of the *BtB* algorithm, including input/output.

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

Table 1. Steps taken to find the exact mRG association rules.

dataset	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
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

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 1. 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.

4 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 this 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. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. In: Proc. of the 3rd Pacific-Asia Conf. on Methodologies for Knowledge Discovery and Data Mining (PAKDD '99), London, UK, Springer-Verlag (1999) 194–203
2. Mansour-Chemaly, M., Haddy, N., Siest, G., Visvikis, S.: Family studies: their role in the evaluation of genetic cardiovascular risk factors. Clin. Chem. Lab. Med. **40**(11) (2002) 1085–1096
3. Weiss, G.: Mining with rarity: a unifying framework. SIGKDD Explor. Newsl. **6**(1) (2004) 7–19
4. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In: Proc. of the 5th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD '99), New York, NY, USA, ACM Press (1999) 337–341
5. Yun, H., Ha, D., Hwang, B., Ryu, K.: Mining association rules on significant rare data using relative support. Journal of Systems and Software **67**(3) (2003) 181–191
6. Koh, Y., Rountree, N.: Finding Sporadic Rules Using Apriori-Inverse. In: Proc. of the 9th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD '05), Hanoi, Vietnam. Volume 3518 of Lecture Notes in Computer Science., Springer (May 2005) 97–106
7. Koh, Y., Rountree, N., O'Keefe, R.: Mining Interesting Imperfectly Sporadic Rules. In: Proc. of the 10th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD '06), Singapore. Volume 3918 of Lecture Notes in Computer Science., Springer (April 2006) 473–482
8. Szathmary, L., Napoli, A., Valtchev, P.: Towards Rare Itemset Mining. In: Proc. of the 19th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI '07). Volume 1., Patras, Greece (Oct 2007) 305–312
9. Wang, K., Jiang, Y., Lakshmanan, L.V.S.: Mining unexpected rules by pushing user dynamics. In: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2003) 246–255
10. Mannila, H., Toivonen, H.: Levelwise Search and Borders of Theories in Knowledge Discovery. Data Mining and Knowledge Discovery **1**(3) (September 1997) 241–258
11. Kryszkiewicz, M.: Concise Representations of Association Rules. In: Proc. of the ESF Exploratory Workshop on Pattern Detection and Discovery. (2002) 92–109
12. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining Frequent Patterns with Counting Inference. SIGKDD Explor. Newsl. **2**(2) (2000) 66–75
13. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-Sets: A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. Data Mining and Knowledge Discovery **7**(1) (Jan 2003) 5–22
14. Calders, T., Rigotti, C., Boulicaut, J.F.: A Survey on Condensed Representations for Frequent Sets. In Boulicaut, J.F., de Raedt, L., Mannila, H., eds.: Constraint-Based Mining. Volume 3848 of LNCS. Springer-Verlag (2005)
15. Boros, E., Gurvich, V., Khachiyan, L., Makino, K.: On the Complexity of Generating Maximal Frequent and Minimal Infrequent Sets. In: Proc. of the 19th Annual Symp. on Theoretical Aspects of Computer Science (STACS '02), London, UK, Springer-Verlag (2002) 133–141
16. Kryszkiewicz, M.: Representative Association Rules. In: Proc. of the 2nd Pacific-Asia Conf. on Research and Development in Knowledge Discovery and Data Mining (PAKDD '98), Melbourne, Australia, Springer-Verlag (1998) 198–209
17. Szathmary, L.: Symbolic Data Mining Methods with the Coron Platform. PhD Thesis in Computer Science, Univ. Henri Poincaré – Nancy 1, France (Nov 2006)