Average Number of Frequent and Closed Patterns in Random Databases

Loïck LHOTE, François RIOULT, Arnaud SOULET

GREYC, CNRS - UMR 6072, Université de Caen
F-14032 Caen cedex France
{prenom.nom}@info.unicaen.fr

Résumé : Frequent and closed patterns are at the core of numerous Knowledge Discovery processes. Their mining is known to be difficult, because of the huge size of the search space, exponentially growing with the number of attributes. Unfortunately, most studies about pattern mining do not address the difficulty of the task, and provide their own algorithm. In this paper, we propose some new results about the *average* number of frequent patterns, by using probabilistic techniques and we extend these results to the number of closed patterns. In a first step, the probabilistic model is simple and far from the real life since the attributes and the objects are considered independent. Nevertheless according to this model, frequency threshold phenomena observed in practice are explained. We also prove that, for a fixed threshold, the number of frequent patterns is asymptotically exponential in the number of attributes and polynomial in the number of objects whereas, for a frequency threshold proportional to the number of objects, the number of frequent and closed patterns is asymptotically polynomial in the number of attributes without depending on the number of objects.

Mots-clés : data mining, average analysis, frequent and closed patterns

1 Introduction

In Knowledge Discovery in Databases, the goal is to find information in databases which describe the *objects* under study with their *attributes*. More precisely, we are trying to find interesting conjunctions of attributes, called *patterns*. These patterns are more or less present in the database, and are qualified by their *frequency* : it is the number of objects containing the pattern. When this quantity rises above a user-defined threshold, the pattern is said *frequent*.

Among others, frequent patterns are at the core of many data mining processes. They give a first piece of information, telling that some conjunctions of attributes are significantly present in the data. They are very useful, e.g. for the association rules discovery, which can ground classification methods. Frequent pattern mining has been well studied, because it is the first stage leading to association rules. Finding these patterns is algorithmically hard, while it is easy to derive association rules from them. In fact, the

search space is exponentially large with the number of attributes, and becomes rapidly intractable.

In this article, we are also interested in *closed* patterns. A closed pattern is the maximal pattern (w.r.t. the inclusion) of the set of patterns having the same frequency and sharing the same attributes. When they are associated with the corresponding pattern of objects containing the pattern of attributes and being also closed, both constitute a *concept*. Conceptual learning is a hot topic (Wille, 1992), and closed patterns is an easy way to non redundant association rules (Zaki, 2000). Their mining has then been widely examined (Kuznetsov & Obiedkov, 2002; Fu & Mephu Nguifo, 2004).

Unfortunately, most studies about pattern mining provide their own solution for solving the mining problem, and sometimes give the complexity of their algorithm, but the theoretical aspects of the difficulty of mining is rarely addressed. The exponential size of the search space is always recalled but only gives an upper bound on the number of frequent patterns, furthermore in the worst case.

In this article, we propose new results about the *average* number of frequent patterns, by using probabilistic techniques. We also give the average number of concepts (or closed patterns, see Section 2) for a frequency threshold proportional to the number of objects. We will see that these results confirm the intuition about the difficulty of the task, by showing that the number of patterns is exponentially large with the number of attributes, and polynomial with the number of objects. Besides, if the frequency threshold is a proportion of the number of objects (10% for example), the average number of frequent patterns is polynomial with the number of attributes, without depending on the number of objects.

The organization of this paper is as follows : we present in Section 2 some definitions and properties about pattern mining in databases, and give the main results of our work in Section 3. We change the model by adding more constraints in Section 4 and end the presentation with some open problems (Section 5). Section 6 is a short conclusion and Appendix A and B gather the proofs of the theorems.

2 Preliminaries

2.1 Notations

A database contains the objects under study, which are described by their attributes. It is usually a boolean matrix, where objects are drawn on the rows, and the binary attributes are the columns. We will not discuss here about the methods for obtaining such a boolean matrix, starting from continuous or multi-valued attributes (see (Srikant & Agrawal, 1996) for an example).
In this article, we will have to distinguish two frameworks :

1. the transactional (consumer bag) framework is the most classical : objects are called transactions and represent a list of purchase. Every bought product is an attribute, and is often called item. It is absent or present in the transaction ;
2. the attribute/value framework is related to the database domain : every continuous attribute is discretized and transformed into several new boolean attributes.

We will yet use the same notation for both frameworks considering that, at the end, we only use boolean attributes. We will come back again to the differences between both frameworks when specifying the probabilistic model (see Section 3.2). The set of attributes is denoted $\mathcal{A} = \{1..m\}$ and the set of objects is $\mathcal{O} = \{1..n\}$. A pattern is a subset of $\mathcal{A}$, and the collection of patterns is denoted by $2^{\mathcal{A}}$.

A database is a subset of $\mathcal{A} \times \mathcal{O}$ and can be represented by a $n \times m$ matrix $(\chi_{i,j})_{i=1..n,j=1..m}$. We can also consider that a database is a set of transactions ; then we will write that a pattern A is supported by a transaction T if $A \subset T$. The *support* of A is the set of all transactions containing A, and the frequency of A is the size of its support. A is said to be γ-frequent if its frequency is over a user-defi ned threshold γ :

Definition 1 (frequent pattern)
Let $\mathcal{B} = (\chi_{i,j})_{i=1..n,j=1..m}$ be a binary database with m items and n transactions and γ a strictly positive integer. A γ-frequent A is a pattern such that $|support(A)| \geq \gamma$.

During the demonstrations, we will use a matrix vision of the support : for all j in A and i in $support(A)$, $\chi_{i,j} = 1$, and for all i in $\mathcal{O} \backslash support(A)$, there exists j in A such that $\chi_{i,j} = 0$.
We now give, in this framework, the defi nition of a γ-closed pattern :

Definition 2 (frequent closed pattern)
Let $\mathcal{B} = (\chi_{i,j})_{i=1..n,j=1..m}$ be a binary database with m items and n transactions and γ a strictly positive integer. A pattern A is γ-closed if :
- *A is γ-frequent pattern,*
- *for all j in $\mathcal{A} \backslash A$, there exists i in $support(A)$ such that $\chi_{i,j} = 0$.*

2.2 Pattern mining

The fi rst and most popular algorithm for mining frequent patterns is A-PRIORI (Agrawal & Srikant, 1994). The key idea is to use the anti-monotonous property of the frequency constraint, which entails that every subset of a frequent pattern is frequent as well, or reciprocally that a superset of an infrequent pattern is infrequent. Starting from frequent items, candidate patterns with two items are built and their frequency is checked in the database. When a candidate is not enough present, its supersets are pruned and will not ground any further candidate. New candidates are produced by joining two frequent patterns having the same prefi x, and again checked in the database, etc.

With this method, patterns are mined with a level-wise strategy, computing them by increasing size. A-PRIORI requires only one database scan to check all candidates at each level : there will be as much database scans as the size of the largest frequent pattern. If we consider that the bottleneck of such a method lies in database accesses, the complexity of A-PRIORI regarding this criteria is good.

The concept of positive and negative border is very useful, in order to more precisely analyze the complexity. The positive border gathers the maximum frequent patterns, with respect to the inclusion order. The negative border offers a dual vision : it brings together the minimum infrequent patterns. Gunopulos et al. have shown that mining

the frequent patterns requires as many database accesses as there are elements in both borders (Gunopulos *et al.*, 1997a).

Since twenty years, closed pattern mining is well studied, but the known methods provide all closed patterns, while we are, in the context of data mining, only interested in the most frequent. Recent works combine both approaches, and the closed patterns can also be mined in a level-wise manner, by using the free (Calders & Goethals, 2003) or key patterns (Pasquier *et al.*, 1999), because they are the generators of the closed patterns.

2.3 Related work

We recall in this section some results about the complexity of frequent pattern mining. As we will see, we are aware about the diffi culty of the mining task : Gunopulos et al. have shown that deciding whether there exists a frequent pattern with t attributes is NP-complete (Gunopulos *et al.*, 1997b; Purdom *et al.*, 2004). The associate counting problem is #P-hard. But we are not really aware about the number of frequent patterns. In fact, as far as we know, there does not exist such results in the literature.

The reason is that the search space is well known to be exponentially large with the number of attributes, and the worst case (e.g. a database where $\chi_{i,j} = 1$ for all i and j, see Figure 1-a) gives $2^m - 1$ frequent patterns (with the minimum frequency threshold $\gamma = 1$). In the middle matrix where $\chi_{i,i} = 0$ (see Figure 1-b), there are $2^m - 2$ closed patterns (with $\gamma = 1$). Finally, in the matrix of the Figure 1-c (Boros *et al.*, 2002), there are k maximal frequent patterns (k is such that $n = k\gamma$), $2^k - 2$ closed patterns, and more than $2^{k(l-1)}$ frequent patterns (l is such that $m = kl$). Of course, it is a pathological example, but we have here a situation where the number of closed patterns is exponentially larger than the number of maximal patterns, and the number of frequent patterns is again exponentially larger than the number of closed patterns.

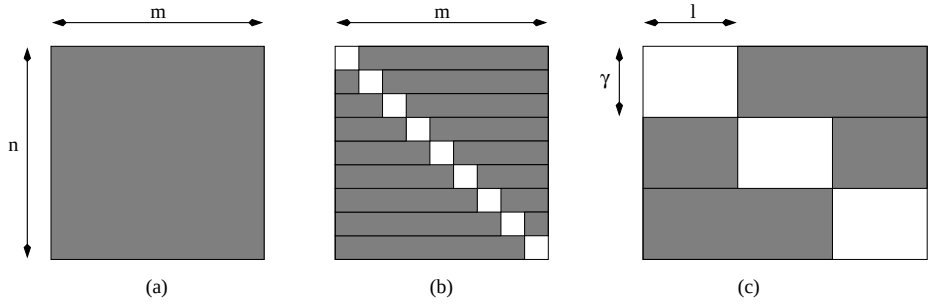

FIG. 1 – Three worst cases for pattern mining (when it is fi lled, it means that there are 1 in the matrix, otherwise there are 0).

Average analysis considerations might then provide interesting results. We found one such study (Purdom *et al.*, 2004), but it is related to the failure rate of A-PRIORI. It is useful for predicting the number of candidates that the algorithm will have to check.

This work confirms the results of (Geerts *et al.*, 2001), who used an upper bound. On other hand, in the seminal paper (Agrawal *et al.*, 1996), the authors of the A-PRIORI algorithm have explained that there are very few long patterns in a random database, and we will reuse the same probabilistic model.

We end this section with quoting (Dexters & Calders, 2004), which gives bounds on the size of the set of k-free patterns (Calders & Goethals, 2003). The authors provide a link between the number of free patterns and the maximum length of such a pattern. Even if this work is hard to relate to ours, we will have to investigate it further.

3 Results with the transactional framework

3.1 Hypothesis

In the following, we are interested in computing the average number of frequent and closed patterns, with respect to a certain minimum frequency threshold γ. All the provided results are asymptotic (i.e. for n and m large) so that the way the frequency threshold is growing with n is important. In practice, two cases are generally distinguished :

Hypothesis 1 (fixed case)
γ is fixed and small when compared to the number n of objects.

For example, γ can be fixed to ten transactions, when there are 100 000 transactions in the database.

Hypothesis 2 (proportional case)
γ is a ratio of n. In this case, we will say that there exists $r \in]0,1[$ such that $\gamma = rn$.

Since we do not have *infinite* databases, the percentage r must not be too small in practice. Nevertheless in our theoretical framework, r can be taken as small as we want but the speed of convergence of our asymptotics decelerates. The distinction between both hypothesis will be useful during the proof of our results : if γ is fixed and small, some approximation can be performed which could not be made if it is a percentage of n. When γ is a ratio of n, some threshold phenomena appears in an integral, which can be exploited by a Laplace's method.

Figure 2 shows the difference between a fixed γ and a variable one. The whole set of (closed) patterns is a lattice where all the patterns with the same cardinality are present on the same horizontal line. Besides, the most general patterns which have the lowest cardinality, are in the top of the lattice. Thus, the lattice may be represented by a rhombus and frequent (closed) patterns correspond to the grey superior part. This figure emphasizes the fact that a variable γ cuts the lattice with preserving the same proportion between both parts. With a fixed γ, this proportion is no longer preserved.

We also have to safely define the ratio between the number of objects and the number of attributes. We therefore require from n and m that they are polynomially linked, i.e. there exists a constant c such that $\log m \sim c \log n$. Let us note that this assumption is only useful for the asymptotic provided in the Theorem 1.

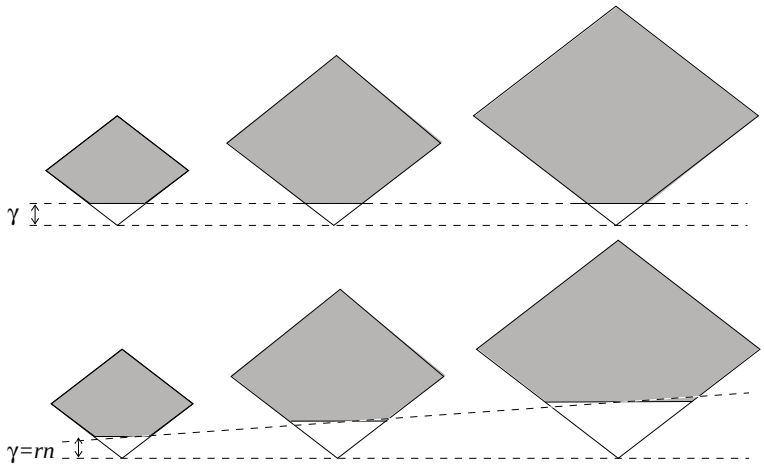

FIG. 2 – Difference between a fixed γ and a variable one (the empty set is at the top of the lattice, and the complete set of attributes is at the bottom. In the case where γ is fixed, the white parts of the lattices, which gather the infrequent patterns, seem to have an equal size, but it is false in practice)

3.2 Probabilistic model

We assume in this section that we are in the transactional framework. The probabilistic model we now describe is very simple. Since we can not appreciate *in advance* the correlations existing in real databases,we will suppose that :
The database $(\chi_{i,j})_{i=1..n,j=1..m}$ *forms an independent family of random variables which follows the same Bernoulli law of parameter* p *in* $]0,1[$.

Figure 3 provides an example of such a transactional database on the left chart : there is no constraint w.r.t. the columns on the number of 1 in each line. This model is far from the reality. Indeed, an equivalent in Information Theory is to modelize the French language with a memoryless source that respects the probability of each letter. The result is not very good but theoretical analysis can be yet lead. In the Section 4, this model is improved in order to handle items coming from continuous or multi-valued attributes. Nevertheless, we will again suppose that the objects are independent.

3.3 Results

This probabilistic model leads to a simple analysis of the average number of γ-closed and γ-frequent patterns. The next theorem sums up our first result for a fixed frequency threshold.

Theorem 1
If the positive integer γ *is fixed (hypothesis 1, Section 3.1) and if there exist a constant* c *such that* $\log m \sim c \log n$*, then for large* n *and* m*, the average number of* γ*-frequent*

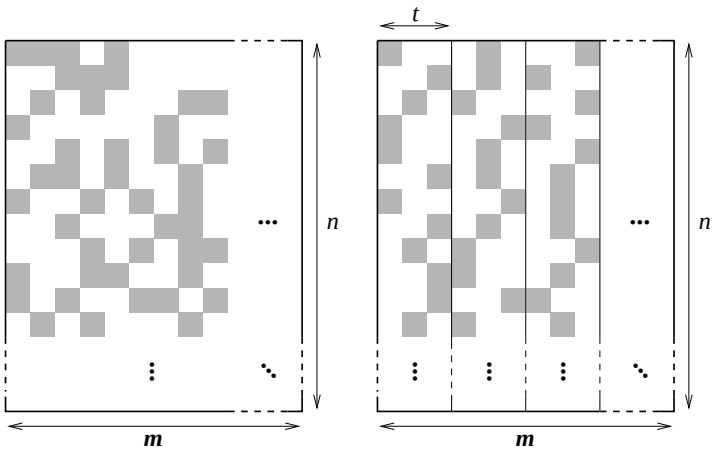

FIG. 3 – Transactional and attribute modelizations of databases (a grey square corresponding to a 1 in the matrix)

patterns $F_{m,n,\gamma}$ satisfies

$$F_{m,n,\gamma} \sim \binom{n}{\gamma}(1+p^\gamma)^m$$

This theorem states that the average number of γ-frequent patterns is asymptotically exponential in the number of attributes and polynomial in the number of objects. This is not really surprising, because we already had this intuition when we studied the search space (which is exponentially large with the number of attributes). Remark nonetheless that the average behavior is far from the worst case, which is 2^m. In addition, the denser the matrix is, the more frequent patterns there are : this is natural. Let us notice that the corresponding proof (see Appendix A) provides the exact asymptotic :

$$F_{m,n,\gamma} = \binom{n}{\gamma}(1+p^\gamma)^m \left[1 + O\left(n\left(\frac{1+p^{\gamma+1}}{1+p^\gamma}\right)^m\right)\right]$$

The following theorem gives a link between the average number of γ-closed patterns and the number of γ-frequent patterns :

Theorem 2
If γ satisfies $\gamma > \lfloor (1+\epsilon)\log m/|\log p| \rfloor$ for an ϵ strictly positive, then the average number of γ-frequent patterns and the average number of γ-closed patterns $C_{m,n,\gamma}$ are equivalent,

$$C_{m,n,\gamma} \sim F_{m,n,\gamma}.$$

We now detail the result of this theorem with the help of the database `T10I4D100K`, which has $n = 100000$ objects, $m = 1000$ attributes, and its density is $p = 0.01$. This dataset is generated by Srikant's synthetic data generator (Agrawal & Srikant, 1994),

and is available on the FIMI website[1]. We used this dataset to illustrate our aim since it has a large number of objects and attributes.

The threshold $\lfloor \log m / |\log p| \rfloor$ for γ involved in the theorem 2 is in practice very low w.r.t. the number of objects. For instance, the theorem applies on `T10I4D100K` when $\gamma > 1.5$. We mined the frequent and the closed patterns in this dataset with Uno's implementations for the FIMI (Uno & Satoh, 2003) and plotted on the Figure 4 the number of patterns, w.r.t. the threshold γ. When γ is greater than 20, we can see that the number of frequent patterns and the number of closed pattern are almost the same.

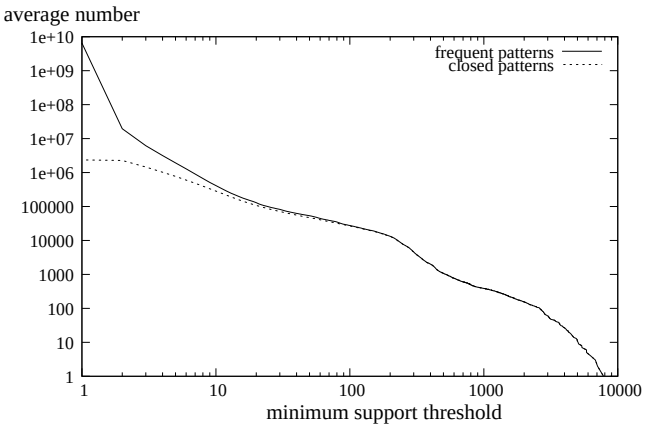

FIG. 4 – Average number of frequent/closed patterns on `T10I4D100K`

It is hard to give an intuition of this surprising result, because it is justified by an approximation which can be realized on the asymptotic (see demonstration on Appendix A). This phenomena can be explained by the poorness of our probabilistic model, which does not handle correlations. Closed patterns are normally useful, because they can summarize correlations. In the conditions of the theorem, almost all the frequent patterns are also closed and A-PRIORI has a better behavior than those algorithms based on the closed patterns.

Now, we consider the average number of patterns with the second hypothesis :

Theorem 3
If γ satisfies $\gamma = \lfloor rn \rfloor$ with $r \in]0,1[$ (hypothesis 2, Section 3.1), then the average number of γ-closed patterns and γ-frequent patterns satisfies for large m and n

$$C_{m,n,\gamma} \sim F_{m,n,\gamma} \sim \binom{m}{j_0}, \quad \textit{where} \quad j_0 = \left\lfloor \frac{\log r}{\log p} \right\rfloor .$$

In other words, j_0 is such that $p^{j_0+1} < r < p^{j_0}$.

[1]Frequent Itemset Mining Implementations is a workshop of the IEEE International Conference on Data Mining (ICDM) http ://fi mi.cs.helsinki.fi /

This theorem is very important, because it states that the average number of frequent patterns (and closed patterns) is **polynomial with the number m of attributes** for a frequency threshold proportional to the number of objects. This is again surprising, because the search space is theoretically exponentially growing with m. Besides, this average number of frequent patterns does not depend on the number n of objects. In the future, we will reuse this result to justify applications of sampling techniques.

4 Results with the attribute/value framework

The preceding results show that our model can be improved. We now try to handle correlations in the data.

In practice, items often come from continuous attributes, that are split. For instance, the attribute *size of the patient* can be split into three items *small, medium, tall*. The previous modelization allowed a patient to be small and tall at the same time, while it is impossible. The new modelization considers these kinds of correlations. Nevertheless, we restrict ourselves to the case where all multi-valued or continuous attributes lead to the same number of boolean attributes $t > 1$. On the right chart, Figure 3 proposes an example of dataset where $t = 3$: there can be only one 1 in each triple of columns.

Since all the original attributes have the same size t, there exists one positive integer m_1 such that the number m of boolean attributes satisfies $m = m_1 t$. The new probabilistic model is based on the following hypothesis :

The database $(\Delta_{i,j} = (\chi_{i,tj+1}, \chi_{i,tj+2}, \ldots, \chi_{i,tj+t}))_{i=1..n, j=0..m_1-1}$ *forms an independent family of random variables with the same uniform law on the set composed with the sequences of size* t *with only one* one *and* $(t-1)$ *zeros (the density of the matrix is* $\frac{1}{t}$*).*

Once more, this model is far from the reality but its equivalent one in Information Theory would be to modelize the French language by a memoryless source that emits trigrams (if $t = 3$) according to their probability. The result is then better than our first modelization.

Using this model, our results are similar to the previous section. The proofs are also similar (see Appendix B).

Theorem 4
If the positive integer γ *is fixed, then the average number of* γ*-frequent patterns* $F_{m,n,\gamma}$ *satisfies for large* m *and* n

$$F_{m,n,\gamma} = \binom{n}{\gamma} \left(1 + \left(\frac{1}{t}\right)^{\gamma-1}\right)^{m_1} \left[1 + O\left(n \left(\frac{1 + (1/t)^{\gamma}}{1 + (1/t)^{\gamma-1}}\right)^{m_1}\right)\right]$$

Theorem 5
If γ *satisfies* $\gamma > \lfloor (1+\epsilon) \log m_1 / \log t \rfloor$ *for an* ϵ *strictly positive, then the average number of* γ*-frequent patterns and the average number of* γ*-closed patterns* $C_{m,n,\gamma}$ *are*

equivalent,

$$C_{m,n,\gamma} \sim F_{m,n,\gamma}.$$

Theorem 6
If γ satisfies $\gamma = \lfloor rn \rfloor$ with $r \in]0,1[$ which is not a power of p, then the average number of γ-closed patterns and γ-frequent patterns satisfies

$$C_{m,n,\gamma} \sim F_{m,n,\gamma} \sim \binom{m_1}{j_0} t^{j_0}, \quad \textit{where} \quad j_0 = \left\lfloor \frac{-\log r}{\log t} \right\rfloor.$$

Theorems 4, 5 and 6 show that the behavior of the asymptotics are very close to those proposed with the former modelization. Nevertheless, the number of γ-frequent patterns with the new model is exponentially lower for fixed γ. Indeed, the factor between the two modelizations is given by

$$\left(\frac{(1+(1/t)^{\gamma-1})^{1/t}}{1+(1/t)^{\gamma}} \right)^{m} = \delta^m \qquad \text{with } \delta < 1$$

It let us think that correlations entail an **exponential decay** on the number of frequent patterns (even if δ is near 1). Thus, this new model really refines the previous results.

Finally with $\gamma = \lfloor rn \rfloor$, the asymptotic is the same than in the first modelization and then, still polynomial.

5 Open problems

We now focus our intention on problems that we did not treat here or manage to solve :

1. What is the average number of γ-closed patterns for fixed γ ? Our feeling is that this number is asymptotically equivalent to the number of closed patterns of size around $\log n / |\log p|$ and frequency around $\log m / |\log p|$.
2. What is the average number of γ-closed/frequent patterns for other function γ such that $\gamma = \sqrt{n}$? The proof of Theorem 3 might be adapted to this context. In particular, the integer j_0 was chosen such $p^{j_0+1} < (\gamma-1)/(n-1) \approx r < p^{j_0}$. By extension, fixing $j_0 = \log_p(\gamma-1)/(n-1)$ we suppose that the number of frequent patterns is $\binom{m}{j_0}$.
3. What is the average size of the biggest frequent pattern ? It corresponds to the number of steps that A-PRIORI Algorithm performs.
4. The positive border is the set of γ-frequent patterns (or equivalently γ-closed ones) whose all supersets are infrequent. What is the average cardinal of the positive border ? This average is given by

$$\sum_{j=1}^{m} \binom{m}{j} \sum_{i=\gamma}^{n} \binom{n}{i} p^{ij} (1-p^j)^{n-i} \left(\sum_{u=0}^{\gamma-1} \binom{i}{u} p^u (1-p)^{i-u} \right)^{m-j}.$$

   For $r > p$, it tends to zero but for $r < p$, the term $(\sum)^{m-j}$ goes from 0 to 1 around $i = pn$. We did not manage to find the asymptotic.

5. The negative border is the set of patterns which are not γ-frequent, and whose all subsets are γ-frequent patterns. What is the average cardinal of the negative border ?

Of course, this list is not exhaustive.

6 Conclusion

In this paper, we gave the average number of frequent or closed patterns in a database, according to the frequency threshold, the number of attributes, objects, and the density of the database. We first used a simple model for the database, consisting in an independent family of Bernouilli random variables. We also provided the results with an improved modelization handling correlations in the attributes.

Our asymptotic results are useful in order to better understand the complexity of the frequent or closed pattern mining task. They explain the efficiency of the frequent pattern mining compared to the closed pattern one on databases close to our models. Furthermore, we emphasized the gap between two choices for the minimal frequency threshold (fixed or not) when the size of pattern lattice grows. In the first case, the average number of patterns is exponential with the number of attributes and polynomial with the number of objects. In the second case, it only polynomially depends on the number of attributes.

In further work, we want to take into account the correlations between objects in order to study the frequent and closed pattern mining on corresponding databases. Besides, we would like to propose a sampling method to estimate the number of patterns starting from a database and a minimum frequency threshold.

A Proofs with the transactional framework

We now prove the theorems. Let us recall that n and m are polynomially linked, i.e. there exist a constant c such that $\log m \sim c \log n$. Thus in the following, both parameters tend to infinity. The first lemma gives simple formulae directly deduced from the definitions for the average number of γ-frequent patterns and γ-concepts. This lemma is sufficient to show the result 2. The second lemma is an integral reformulation of a part of the previous formulae. The third lemma gives asymptotics for the integral part. Finally, we prove the theorem.

Lemma 1
The average number of γ-frequent patterns satisfies

$$F_{m,n,\gamma} = \sum_{j=1}^{m} \binom{m}{j} \sum_{i=\gamma}^{n} \binom{n}{i} p^{ij} (1-p^j)^{n-i}.$$

The average number of γ-concepts satisfies

$$C_{m,n,\gamma} = \sum_{j=1}^{m} \binom{m}{j} \sum_{i=\gamma}^{n} \binom{n}{i} p^{ij} (1-p^i)^{m-j} (1-p^j)^{n-i}.$$

Proof 1 (Lemma 1)
Fix (A, O) a γ-frequent pattern. The cardinal of a set E is noted $|E|$. Since O is the support of A, for all index in $A \times O$, there is a one in the matrix. But the probability o having a one is p so that the probability of having a one at each index of $A \times O$ is $p^{|A||O|}$. In addition, O is the greatest set containing all the items of A, so that for all transactions in $\mathcal{O}\backslash O$, there is at least one zero at an index of A. The probability of satisfying this last condition is $(1 - p^{|A|})^{n-|O|}$.

If (A, O) is a concept, the probability that A is the greatest set containing O is by symmetry $(1 - p^{|O|})^{m-|A|}$. Now, summing over all the possible cardinalities for A and O, we get both formulae.

Proof 2 (Theorem 2)
Both formulae are sufficient to prove Theorem 2. Indeed, if γ satisfies $\gamma > \lfloor (1 + \epsilon) \log m / |\log p| \rfloor$ for an ϵ strictly positive, then

$$\begin{aligned}(1-p^i)^{m-j} &\leq (1-p^{(1+\epsilon)\log m/|\log p|-1})^{m-j} \\ &= (1-\frac{1}{pm^{1+\epsilon}})^{m-j} \\ &\to 1.\end{aligned}$$

Theorem 2 follows from this equivalence.

The next lemma expresses the sum over i in $F_{m,n,\gamma}$ with an integral. This is the key point of all the proofs, since the way we approximate the integral leads to two different asymptotics for γ fi xed or linear.

Lemma 2
One has the integral equality :

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \gamma \binom{n}{\gamma} \int_0^x t^{\gamma-1}(1-t)^{n-\gamma} dt.$$

Proof 3 (Lemma 2)
Expanding $(1 - x)^{n-i}$ leads to

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{i=\gamma}^{n} \binom{n}{i} \sum_{u=0}^{n-i} \binom{n-i}{u} (-1)^u x^{i+u}.$$

Now, the change of variable $v = u + i$ and the inversion of both signs sum gives the new equality

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{v=\gamma}^{n} \binom{n}{v} x^v \sum_{i=\gamma}^{v} \binom{v}{i} (-1)^{v-i}.$$

A simple induction shows that the second sum simplifies into

$$(-1)^{v-\gamma} \binom{v-1}{\gamma-1}.$$

Hence, the previous inequality becomes

$$\sum_{i=\gamma}^{n}\binom{n}{i}x^i(1-x)^{n-i}=\sum_{v=\gamma}^{n}\binom{n}{v}x^v(-1)^{v-\gamma}\binom{v-1}{\gamma-1}.$$

Now, the binomials simplify, $\binom{n}{v}\binom{v-1}{\gamma-1}=\frac{\gamma}{v}\binom{n}{\gamma}\binom{n-\gamma}{v-\gamma}$ *and the change of variable* $w=v-\gamma$ *gives the new expression*

$$\sum_{i=\gamma}^{n}\binom{n}{i}x^i(1-x)^{n-i}=\gamma\binom{n}{\gamma}\sum_{w=0}^{n-\gamma}\binom{n-\gamma}{w}(-1)^w\frac{x^{w+\gamma}}{w+\gamma}.$$

To conclude, remark that the second sum is zero when $x=0$ *and that the derivative according to* x *is exactly* $x^{\gamma-1}(1-x)^{n-\gamma}$*. The lemma follows.*

We can now prove Theorem 1.

Proof 4 (Theorem 1)
Let f *be the function* $f(x)=(1-x)^{n-\gamma}$*. The sign of the derivatives of* f *alternates so that, the Taylor expansion of* f *entails the bounds,*

$$\sum_{l=0}^{2k+1}\frac{f^{(l)}(0)}{l!}x^l\leq f(x)\leq\sum_{l=0}^{2k}\frac{f^{(l)}(0)}{l!}x^l$$

for all positive integer k*. Now the derivatives satisfy* $f^{(l)}(0)=(-1)^l(n-\gamma)\dots(n-\gamma-l+1)x^{n-\gamma-l}$*. A bound of the integral formula is then*

$$\begin{aligned}\int_0^x t^{\gamma-1}(1-t)^{n-\gamma}dt &\approx \int_0^x\sum_{l=0}^{2k+1}\frac{f^{(l)}(0)}{l!}t^{l+\gamma-1}dt\\ &= \sum_{l=0}^{2k+1}\binom{n-\gamma}{l}(-1)^l\frac{x^{l+\gamma}}{l+\gamma}.\end{aligned}$$

Applying Lemma 2 with $F_{m,n,\gamma}$*, using the previous bounds and summing over* j *in* $F_{m,n,\gamma}$ *finally gives*

$$F_{m,n,\gamma}\approx\gamma\binom{n}{\gamma}\sum_{l=0}^{2k+1}\binom{n-\gamma}{l}(-1)^l\frac{(1+p^{l+\gamma})^m-1}{l+\gamma}.$$

In particular for $k=0$*, one has*

$$\binom{n}{\gamma}((1+p^\gamma)^m-1)-\gamma\binom{n}{\gamma}\binom{n-\gamma}{1}\frac{(1+p^{1+\gamma})^m-1}{1+\gamma}\leq F_{m,n,\gamma}\leq\binom{n}{\gamma}(1+p^\gamma)^m-1.$$

To conclude, the condition $\log m\sim c\log n$ *entails that the binomials are polynomial in* m *and it is negligible compared to the exponential part. This finishes the proof of Theorem* 1.

The last lemma describes the asymptotic of the integral when n is large.

Lemma 3
Suppose that γ satisfies $\gamma = \lfloor rn \rfloor$ with r a non-power of p.
For $x > r$,

$$\int_0^x t^{\gamma-1}(1-t)^{n-\gamma}dt = \frac{1}{\gamma\binom{n}{\gamma}}(1+\epsilon_n(x)),$$

with $(\epsilon_n)_n$ a sequence of decreasing functions that converges uniformly to zero.
For $x < r$,

$$\int_0^x t^{\gamma-1}(1-t)^{n-\gamma}dt = \frac{\exp(ng_n(x))}{ng_n'(x)}(1+\widetilde{\epsilon}_n(x)),$$

with $(\widetilde{\epsilon}_n)_n$ a sequence of increasing functions that converges uniformly to zero and

$$g_n(x) = \frac{\gamma-1}{n}\log x + \frac{n-\gamma}{n}\log(1-x).$$

Proof 5 (Lemma 3)
This lemma is the well known Laplace Method. The proof is then let to the reader.

We finally prove Theorem 3.

Proof 6 (Theorem 3)
Integer j_0 is (asymptotically) the lowest integer j such that $p^j > r$. By Lemma 1 and Lemma 2, the average number of frequent patterns satisfies

$$F_{m,n,\gamma} = \sum_{j=1}^{n}\binom{m}{j}\gamma\binom{n}{\gamma}\int_0^{p^j} t^{\gamma-1}(1-t)^{n-\gamma}dt.$$

The sum is then split into two sums $\sum_{j=1}^{j_0} + \sum_{j=j_0+1}^{m}$ and the use of lemma 3 provides the equivalence

$$F_{m,n,\gamma} \sim \sum_{j=1}^{j_0}\binom{m}{j} + \sum_{j=j_0+1}^{m}\binom{m}{j}\gamma\binom{n}{\gamma}\frac{\exp(ng_n(p^j))}{ng_n'(p^j)}.$$

The first sum is equivalent to $\binom{m}{j_0}$ since j_0 is constant. A simple upper bound gives the inequality for the second sum,

$$\sum_{j=j_0+1}^{m}\binom{m}{j}\frac{\exp(ng_n(p^j))}{ng_n'(p^j)} \leq \frac{1}{\gamma-1-p^{j_0}(n-1)}\sum_{j=j_0+1}^{m}\binom{m}{j}p^{\gamma j}(1-p^j)^{n-\gamma+1}.$$

Now, an equivalent of the right sum is

$$\sum_{j=j_0+1}^{m}\binom{m}{j}p^{\gamma j}(1-p^j)^{n-\gamma+1} \sim \binom{m}{j_0+1}p^{\gamma(j_0+1)}(1-p^{j_0+1})^{n-\gamma+1}. \quad (1)$$

Indeed, let $w_j = \binom{m}{j} p^{\gamma j}(1-p^j)^{n-\gamma+1}$. *The ratio* w_{j+1}/w_j *is decreasing with* j *and the ratio* w_{j_0+2}/w_{j_0+1} *satisfies*

$$\frac{w_{j_0+2}}{w_{j_0+1}} = \frac{m-j_0-2}{j_0+3}\exp(n\theta(\gamma,n,p,j_0))$$

$$\textit{with} \quad \theta(\gamma,n,p,j_0) = \frac{\gamma}{n}\log p + \frac{n-\gamma+1}{n}\log(1+p^{j_0+1}\frac{1-p}{1-p^{j_0+1}}).$$

Using that γ/n *tends to* r *as* n *tends to infinity and that* $p^{j_0+1} < r$, *the function* θ *is shown to converge to a strictly negative constant. Hence, the ratio* w_{j_0+2}/w_{j_0+1} *tends to zero as* n *tends to infinity what is sufficient to prove the equivalent of Formula (1). The Stirling formula applied with the binomial* $\binom{n}{\gamma}$ *entails the equivalent*

$$\binom{n}{\gamma} p^{(j_0+1)\gamma}(1-p^{j_0+1})^{n-\gamma} \sim \sqrt{\frac{1}{2\pi r(1-r)n}}\left(\frac{p^{j_0+1}n}{\gamma}\right)^{\gamma}\left(\frac{1-p^{j_0+1}}{1-(\gamma/n)}\right)^{n-\gamma}$$

$$= \exp(n\widetilde{\theta}(\gamma,n,p,j_0)),$$

$$\textit{where} \qquad \widetilde{\theta}(\gamma,n,p,j_0) = \frac{\gamma}{n}\log\frac{p^{j_0+1}n}{\gamma} + (n-\gamma)\log\frac{1-p^{j_0+1}}{1-(\gamma/n)}.$$

Finally, since for all positive x, $\log x \leq x-1$ *and* $\log 1+x \leq x$, *the function* $\widetilde{\theta}$ *is proved to converge to a strictly negative number. It follows that*

$$\gamma\binom{n}{\gamma}\sum_{j=j_0+1}^{m}\binom{m}{j}\frac{\exp(ng_n(p^j))}{ng_n'(p^j)} \to 0.$$

which finishes the proof of Theorem 3.

B Proofs with the attribute/value framework

The proofs are exactly identical. The only change is the first formula for the average number of frequent patterns or concepts.

Lemma 4
The average number of γ*-frequent patterns satisfies*

$$F_{m,n,\gamma} = \sum_{j=1}^{m_1} t^j \binom{m_1}{j}\sum_{i=\gamma}^{n}\binom{n}{i}\left(\frac{1}{t}\right)^{ij}(1-\left(\frac{1}{t}\right)^j)^{n-i}.$$

The average number of γ*-closed patterns satisfies*

$$C_{m,n,\gamma} = \sum_{j=1}^{m_1} \binom{m_1}{j}\sum_{i=\gamma}^{n}\binom{n}{i}\left(\frac{1}{t}\right)^{ij}(1-\left(\frac{1}{t}\right)^i)^{m_1-j}(1-\left(\frac{1}{t}\right)^j)^{n-i}.$$

All the previous proofs extend to these formulae.

Références

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & VERKAMO A. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*.

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules. In *Intl. Conference on Very Large Data Bases (VLDB'94), Santiago de Chile*.

BOROS E., GURVICH V., KHACHIYAN L. & MAKINO K. (2002). On the complexity of generating maximal frequent and minimal infrequent sets. In *Symposium on Theoretical Aspects of Computer Science*, p. 133–141.

CALDERS T. & GOETHALS B. (2003). Minimal k-free representations of frequent sets. In *Proceedings of PKDD'03*.

DEXTERS N. & CALDERS T. (2004). Theoretical bounds on the size of condensed representations. In *ECML-PKDD 2004 Workshop on Knowledge Discovery in Inductive Databases (KDID)*.

FU H. & MEPHU NGUIFO E. (2004). Etude et conception d'algorithmes de génération de concepts formels. *Revue des sciences et technologies de l'information, série ingénierie des systèmes d'information (RSTI-ISI)*, **9**, 109–132.

GEERTS F., GOETHALS B. & VAN DEN BUSSCHE J. (2001). A tight upper bound on the number of candidate patterns. In *Proceedings of ICDM'01*, p. 155–162.

GUNOPULOS D., MANNILA H., KHARDON R. & TOIVONEN H. (1997a). Data mining, hypergraph transversals, and machine learning. In *PODS 1997*, p. 209–216.

GUNOPULOS D., MANNILA H. & SALUJA S. (1997b). Discovering all most specific sentences by randomized algorithms. In *ICDT*, p. 215–229.

KUZNETSOV S. O. & OBIEDKOV S. A. (2002). Comparing performance of algorithms for generating concept lattices. *J. Exp. Theor. Artif. Intell.*, **14(2-3)**, 189–216.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, **24**(1), 25–46.

PURDOM P. W., VAN GUCHT D. & GROTH D. P. (2004). Average-case performance of the apriori algorithm. *SIAM Journal on Computing*, **33**(5), 1223–1260.

SRIKANT R. & AGRAWAL R. (1996). Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, p. 1–12 : ACM Press.

UNO T. & SATOH K. (2003). LCM : An efficient algorithm for enumerating frequent closed item sets. In *Workshop on Frequent Itemset Mining Implementations (ICDM'03)*.

WILLE R. (1992). Concept lattices and conceptual knowledge systems. In *Computer mathematic applied, 23(6-9) :493-515*.

ZAKI M. J. (2000). Generating non-redundant association rules. In *SIGKDD'00, Boston*, p. 34–43.